nature medicine

ARTICLES

https://doi.org/10.1038/s41591-019-0711-0

Autism risk in offspring can be assessed through quantification of male sperm mosaicism

Martin W. Breuss[1,2], Danny Antaki[3,4,5,6], Renee D. George[1,2], Morgan Kleiber[3,4,5], Kiely N. James[1,2], Laurel L. Ball[1,2], Oanh Hong[3,4,5,6], Ileena Mitra[7,8], Xiaoxu Yang[1,2], Sara A. Wirth[1,2], Jing Gu[1,2], Camila A. B. Garcia[1,2], Madhusudan Gujral[3,4,5,6], William M. Brandler[3,4,5,6], Damir Musaev[1,2], An Nguyen[1,2], Jennifer McEvoy-Venneri[1,2], Renatta Knox[1,2,9], Evan Sticca[1,2], Martha Cristina Cancino Botello[10], Javiera Uribe Fenner[10], Maria Cárcel Pérez[11], Maria Arranz[11], Andrea B. Moffitt[12], Zihua Wang[12], Amaia Hervás[13], Orrin Devinsky[14], Melissa Gymrek[7,8], Jonathan Sebat[3,4,5,6*] and Joseph G. Gleeson[1,2*]

De novo mutations arising on the paternal chromosome make the largest known contribution to autism risk, and correlate with paternal age at the time of conception. The recurrence risk for autism spectrum disorders is substantial, leading many families to decline future pregnancies, but the potential impact of assessing parental gonadal mosaicism has not been considered. We measured sperm mosaicism using deep-whole-genome sequencing, for variants both present in an offspring and evident only in father's sperm, and identified single-nucleotide, structural and short tandem-repeat variants. We found that mosaicism quantification can stratify autism spectrum disorders recurrence risk due to de novo mutations into a vast majority with near 0% recurrence and a small fraction with a substantially higher and quantifiable risk, and we identify novel mosaic variants at risk for transmission to a future offspring. This suggests, therefore, that genetic counseling would benefit from the addition of sperm mosaicism assessment.

Clinicians are facing an ever-increasing incidence of autism spectrum disorders (ASD) in the population, without effective strategies available to prevent disease or counsel families. Recent studies have identified gene-damaging de novo mutations (DNMs) in at least 10–30% of simplex ASD cases[1–4], along with the realization that the number of DNMs increases as a function of paternal age at the time of conception, doubling in DNM number in an offspring every 16.5 years of the father's age at the time of conception[5,6]. A DNM, defined as a genetic variant present in an offspring but not detectable in either parent, can have any of several different origins[7,8]. While classically considered as occurring in the fertilized egg at the one-cell stage, most probably occur either postzygotically in the offspring or in a parent, either in the gonads or broadly in a mosaic pattern[9]. DNMs that occur during embryogenesis of a parent cause mosaicism in the soma, the gonads or both, and remain throughout life yet may be undetectable or barely detectable in blood[10]. However, the balance of gonadal-specific compared to broadly distributed DNMs in the father has not been carefully assessed, and thus the role of gonadal mosaicism in DNM recurrence risk remains uncertain.

Knowledge of the rates and mechanisms by which gonadal mutations arise has been advanced through assessment of multiple transmissions of DNMs within families, where approximately 1.3% of DNMs are shared by siblings[11]. Although only 3.8% of offspring DNMs are detectably mosaic in parental blood, this increases to 57.2% if shared by two or more offspring[10,11]. Counterintuitively, DNM recurrence risk decreases by 1.8–2.3% per year of parent age, due to an increase in aging-associated DNMs[10,11], thereby decreasing the relative contribution of parental mosaic variants to mutation burden.

Results

Sperm sequencing allows stratification of variants into low and high recurrence risk. We recruited eight families from our ASD cohort[12,13], where each father agreed to submit a sperm sample for sequencing (Supplementary Dataset 1). Employing 30× whole-genome sequencing (WGS) from blood[12,13], we defined 912 de novo single-nucleotide variants (dSNVs) in the 14 offspring (Fig. 1a and Methods). We then isolated sperm from the ejaculates and performed 200× WGS on paternal blood and sperm cells to determine which dSNVs were detectable in sperm based on three or more mutant reads (Extended Data Fig. 1 and Methods). We found 23 (2.5%) dSNVs that were also detected in paternal blood or sperm, leaving 889 (97.5%) dSNVs undetectable (Fig. 1b

[1]Department of Neurosciences, Howard Hughes Medical Institute, University of California, San Diego, La Jolla, CA, USA. [2]Rady Children's Institute for Genomic Medicine, San Diego, CA, USA. [3]Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, CA, USA. [4]Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. [5]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. [6]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [7]Department of Medicine, University of California, San Diego, La Jolla, CA, USA. [8]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [9]Department of Child Neurology, Weill Cornell Medical College, New York, NY, USA. [10]Child and Adolescent Mental Health Unit, Hospital Universitari Mútua de Terrassa, Barcelona, Spain. [11]Fundació Docència i Recerca Mútua Terrassa, Barcelona, Spain. [12]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY, USA. [13]Research Laboratory Unit, Fundació Docencia i Recerca Mútua Terrassa, Barcelona, Spain. [14]Department of Neurology, Epilepsy Division, New York University School of Medicine, New York, NY, USA. *e-mail: jsebat@ucsd.edu; jogleeson@ucsd.edu

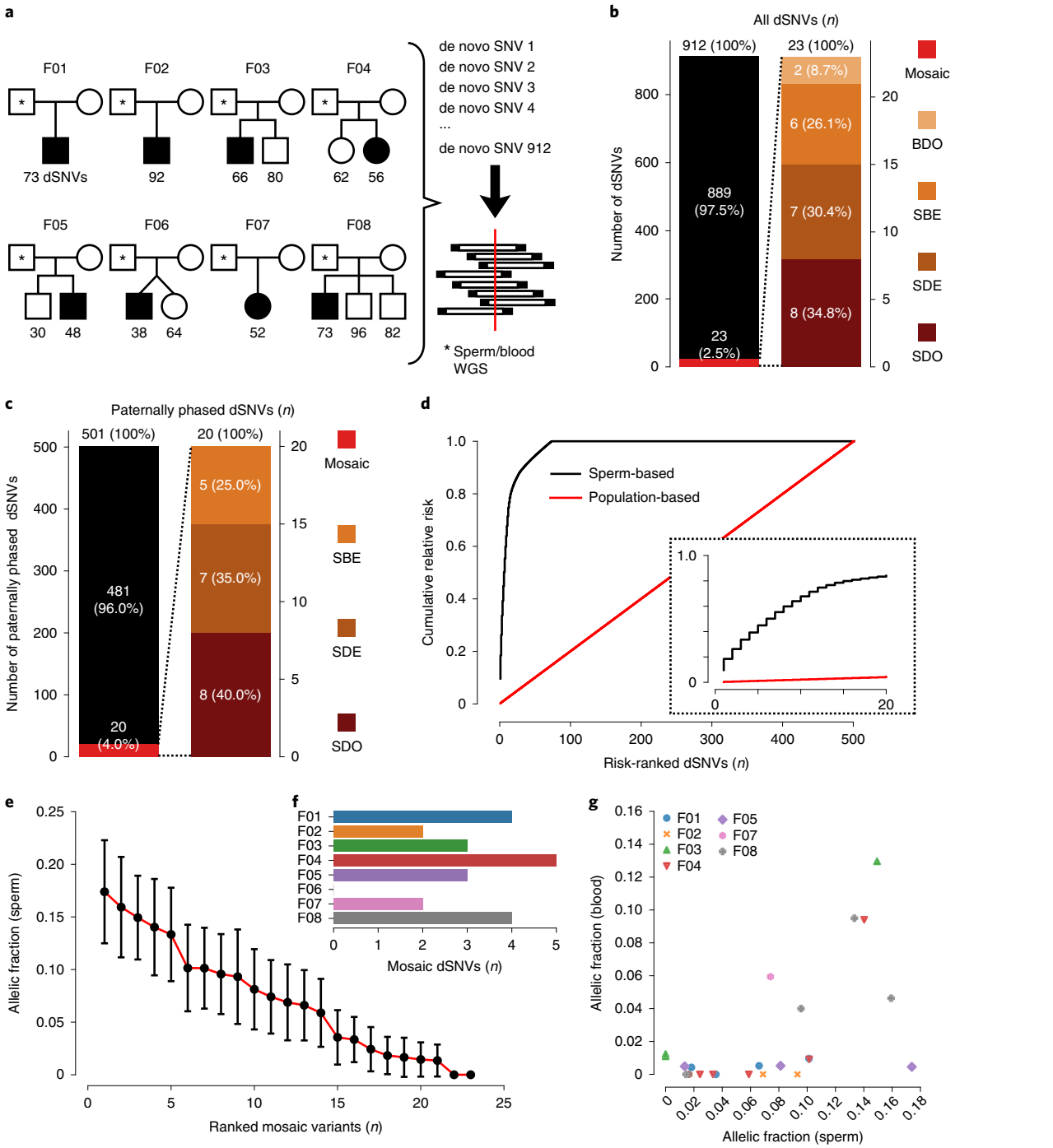


Fig. 1 | Recurrence risk stratification and mosaicism rates of 912 dSNVs are different in sperm compared with blood. a, Eight nuclear families on which 200× WGS analyses of father's sperm and blood were based. dSNVs in offspring were evaluated in paternal sperm and blood using WGS data. Filled symbols, ASD diagnosis. **b**, dSNV assessment in eight families identified 912 dSNVs, of which 23 (2.5%) were detected in father's sperm or blood with ≥3 mutant reads: 34.8% of these were SDO, 30.4% were SDE ($\alpha > 3$), 26.1% were present at equal AF in sperm and blood (SBE) and 8.7% were BDO. **c**, Relative number of paternally phased dSNVs showing evidence (≥3 reads) of mosaicism in blood, sperm or both. **d**, Contribution to cumulative relative recurrence risk for all paternally phased dSNVs. Risk derived from sperm mosaicism (≥1 alternate read, black), assuming equal risk for all variants (red). Dashed box shows only the first 20 identified paternally phased mosaic variants. **e**, Ranked plot of estimated sperm AF (estimated fraction ± binomial 95% CI, based on the fraction of mutant reads; see Supplementary Dataset 2) for all mosaic variants. **f**, Number of mosaic variants found in each father's sperm. F04 had the most, at five, and F06 had none detected. **g**, Sperm versus blood AF for all detected mosaic variants, coded by family. Most sperm mosaic AFs <8% were either SDO or SDE, whereas most mosaic variants >8% were also detected in blood at similar AFs.

and Supplementary Dataset 2). Orthogonal validation of a subset with ultra-deep target amplicon sequencing (TAS) showed a validation rate of ~83% (15/18; Extended Data Fig. 2, Supplementary Dataset 3 and 4 and Methods). All three nonvalidated variants were at allelic fractions (AFs) <3% or located within repetitive elements (SINE or LINE).

Using the ratio of mutant to reference reads in blood and sperm, we defined four dSNV classes: sperm-detectable only (SDO); sperm-detectable enriched (SDE)—for which the AF was >3-fold higher in sperm than in blood (i.e., $\alpha > 3$); sperm–blood equal (SBE, enrichment <3-fold); and blood-detectable only (BDO). Of the 23 variants, 34.8, 30.4, 26.1 and 8.7% were SDO, SDE, SBE and BDO, respectively. Nanopore long-read sequencing of the children allowed phasing of 501 of the 912 dSNVs to the paternal haplotype. Of the 23 mosaic variants, 20 resided on the paternal chromosome (40% SDO, 35.0% SDE, 25.0% SBE and 0% BDO; Fig. 1c). Thus, assessment of blood or using population risk underestimates paternal gonadal mosaicism (PGM) for most mosaic dSNVs (Fig. 1d). Furthermore, most dSNVs are not present in paternal sperm at this sensitivity level and thus have little measurable likelihood of recurrence.

The PGM burden was roughly equally distributed among the eight families (0–5 PGM variants/male), with AFs varying from 17% to the lower detection limit of 1.3% (Fig. 1e). Neither the number of mosaic variants nor their AF correlated with paternal age (Extended Data Fig. 3). We observed a mutational signature for PGM variants consistent with a developmental origin, not observed for nonPGM DNMs (for example, relative decrease in T>C variants[6,10]) (Extended Data Fig. 4). While PGM variants >7% AF were often also detectable in blood (10/11 SBE or SDE), variants below this level were typically restricted to sperm (7/12 SDO). Together, these data are consistent with an origin of PGM during embryonic development of the father, with those occurring earlier showing broader tissue distribution and higher AFs[14].We next assessed the potential of sperm/blood sequencing to measure PGM for de novo structural variants (dSVs) and de novo short tandem-repeat variants (dSTRΔs; see Methods). Among the eight families, F01 had two de novo deletions (dDels) and F06 had one de novo duplication (dDup; Fig. 2a). One of these variants was detectably mosaic in paternal sperm, with an AF of 2–6% (Fig. 2b–d and Extended Data Fig. 5a–d). Among the eight families we identified 126 different dSTRΔs, of which 15 (11.9%) were mosaic (Fig. 2e–j, Extended Data Fig. 5e–h and Supplementary Dataset 5). Because 12 of 15 variants were SDO or SDE, recurrence risk assessment from blood alone would be erroneous for 80%.

PGM extends to ASD pathogenic variants. We next assessed whether clinically pathogenic DNMs could be detected in parental sperm, which could impact clinical decision making. We assessed a cohort of 14 families in which an offspring had ASD attributed to a dSNV or a 1-base pair (bp) dDel based on American College of Medical Genetics guidelines (Fig. 3a, Supplementary Text, Supplementary Table 1 and Supplementary Dataset 1). Using Droplet Digital PCR (ddPCR), three of 14 (21.4%) DNMs were detected as mosaic in sperm, with AFs of 14.47% (F09), 0.56% (F10) and 8.09% (F13) (Fig. 3b, Extended Data Fig. 6a–d and Supplementary Dataset 3). We were successful in phasing the 14.47% and 0.56% AF variants to the paternal haplotype (Supplementary Dataset 2 and 6). Three variants phased to the maternal haplotype posed no risk of PGM, but seven could not be phased, including the 8.09% AF variant. The F13 variant was absent in paternal blood (SDO), and the F09 variant was substantially reduced (SDE) (Extended Data Fig. 7a–c). These results, while representing a small number of DNMs, suggest that a substantial fraction of both paternally phased and unphased disease-related DNMs are detectable as PGM, and thus recurrence risk can be estimated directly.

Two variants showed sperm AF that predicted substantially elevated recurrence above the basal 1% recurrence risk in families (F09 at 14.47% AF and F13 at 8.09% AF)[7]. While F13 had a single child, F09, with a c.1007+1 G>A known pathogenic variant in *GRIN2A*[15,16] (Fig. 3c), had two older siblings lacking criteria for ASD, but deeper questioning revealed that both siblings showed neurodevelopmental abnormalities with no known cause (Fig. 3d and Supplementary Table 2). The middle child showed ADHD and speech impairment and the oldest child had ADHD and seizures, all consistent with *GRIN2A* haploinsufficiency. We collected DNA samples from the whole family and found that the *GRIN2A* c.1007+1G>A variant was heterozygous in all three children (Extended Data Fig. 6e). Thus, the mosaic variant in the father's sperm at 14.47% was transmitted to all three offspring, an unlikely but confirmed event, resulting in pleiotropic clinical features.

In our larger ASD cohort, five families had dSVs detected with standard WGS that were considered risk alleles (Supplementary Text)[12,13]. These included the de novo 22q12.3 dDel in F01 and the 1p36.32 dDup in F06 (Fig. 2a), as well as F18 with a 7q11.23 dDup and a 16p13.11 dDel, F19 with a 15q13.1-q13.3 dDel and F20 with a 10q21.3-q22.1 dDup (Fig. 3e). Several of the CNVs (for example, 15q13.1-q13.3) were flanked by directly oriented segmental duplications, suggesting that they may have arisen during meiosis through nonallelic homologous recombination[17,18]. A meiotic origin of these variants would preclude any possibility of PGM; however, as nonallelic homologous recombination may also occur during mitosis, these were still included in this analysis[19].

We phased all of these variants and found that all bar one, the 7q11.23 dDup, phased to the paternal haplotype. Probe sets were designed to interrogate these variants from sperm using PCR and ddPCR copy number assessment (Fig. 3f and Supplementary Dataset 7). These assays confirmed the presence of the dSV in all tested probands, but did not reveal sperm mosaicism in any additional cases beyond F01 (Fig. 3g–i and Extended Data Fig. 7e–h). The 22q12.3 variant in F01 was mosaic in the father's sperm sample, based on the presence of a junction fragment matching the band in the proband and assessment of the deletion by nested PCR (Fig. 3g–h). ddPCR quantification showed 0.9382 mutant allele abundance in the proband (that is, heterozygous), whereas the father's sperm showed a 0.1538 abundance and his blood 0.0023 abundance (Fig. 3i), suggesting that ~7–8% of sperm carries the deletion. Thus, one of five dSVs was detectably mosaic in parental sperm at an AF that could be considered clinically significant, since this would increase recurrence risk by ~7–8-fold (Extended Data Fig. 7e–h). The specificity of these assays precluded the confident exclusion of mosaicism in paternal sperm except for one additional variant (F20 with a 10q21.3-q22.1 dDup), and thus a negative predictive value is more difficult to calculate for most dSVs.

For three of the four pathogenic variants that were mosaic in sperm, a second semen sample, collected 1–4 months after the first, was subjected to mosaicism analysis by ddPCR (Extended Data Fig. 7d). While all three tested variants were detected at similar AF, the *NR2F1* mutation exhibited a slight, but significant, difference between the two samples ($P < 0.001$). This suggests that mosaicism at these higher AFs is relatively stable over time.

Unbiased analysis of sperm mosaicism detects 9–23 mosaic variants in sperm. We next assessed the value in identifying PGM for variants not yet observed in children. Using the 200× sperm WGS on the eight fathers, we identified mosaic variants using the intersection of variants of MutTect 2 and Strelka 2 (ref. [20,21]), both optimized for mosaic variant detection in one tissue compared to another, as well as MosaicHunter[22], optimized for mosaic variant detection shared between two tissues (Fig. 4a, Extended Data Fig. 8 and Supplementary Dataset 8). Combined, these methods identified 6/23 DNMs (from Fig. 1b) as PGM, since many of these occurred in repetitive sequences that were masked by these callers. This low recall rate was partially due to optimization of the pipeline for specificity (TAS, ~90% validation rate; Extended Data Fig. 9). To increase power for subsequent analyses on variants detected in blood and sperm, we defined three major groups of mosaic mutations—SDO, BSS (blood/sperm shared; includes SDE, SBE and blood detectable

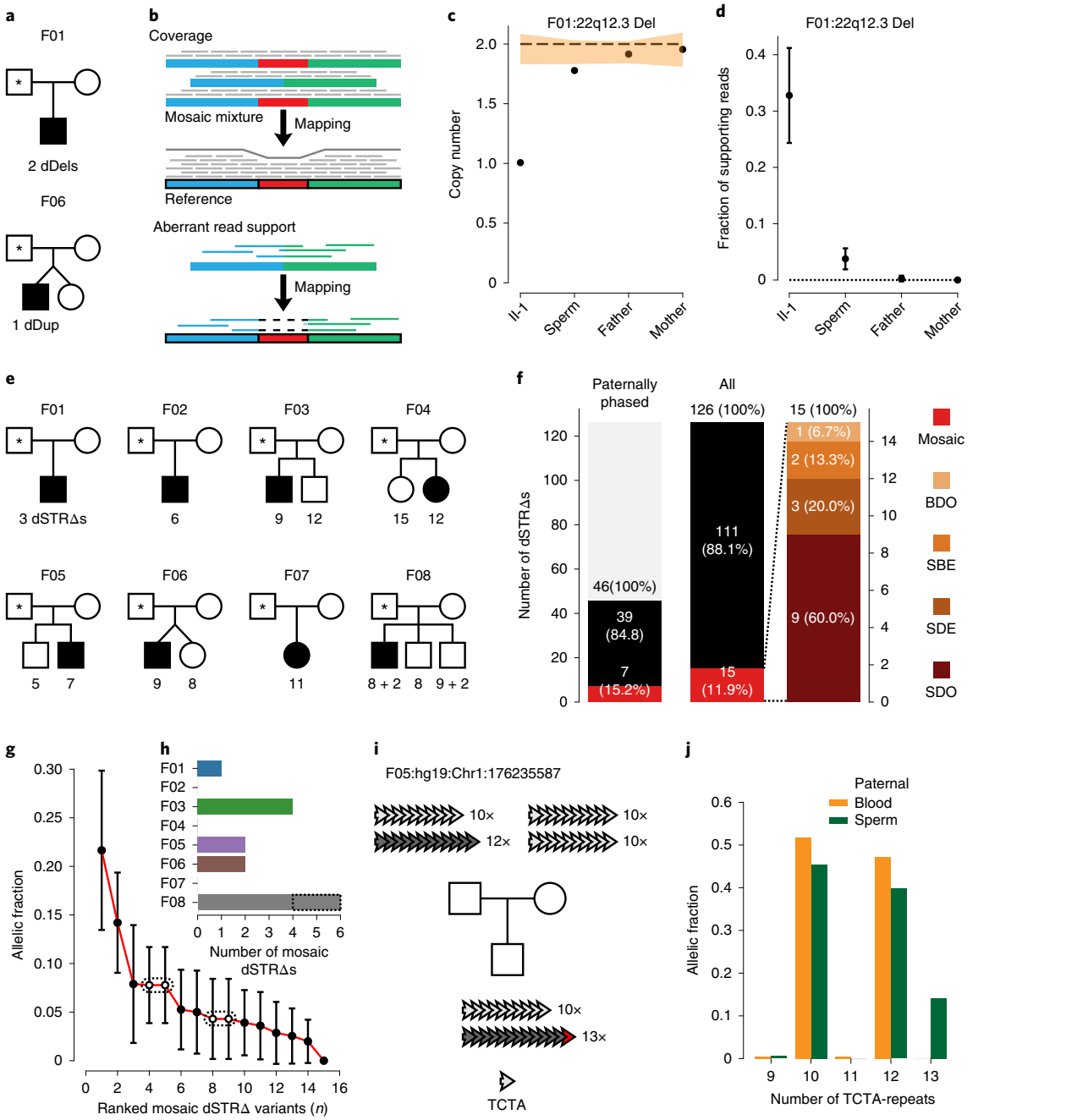


Fig. 2 | Risk stratification can be applied to other classes of DNMs. a, Pedigrees for F01 and F06 and detected dSVs. **b**, Approaches to detection of dSV gonadal mosaicism: coverage and aberrant read support. **c**, Calculated copy number for the 22q12.3 deletion in F01. Dashed line denotes expected copy number (two copies). Orange band denotes ±1 s.d. of the copy number using similarly sized regions across the genome ($n = 1{,}000$ random regions, see Methods). **d**, Estimated fraction of supporting reads (estimated fraction ± binomial 95% CI, based on the fraction of mutant reads; see Supplementary Dataset 7) for the 22q12.3 deletion in F01. **e**, Eight nuclear families and the detected dSTRΔs for each child ($n = 126$ variants, two of which are recurrent in F08). **f**, Gonadal mosaicism assessment for 126 dSTRΔs in father's sperm from eight families. **g**, Ranked plot of estimated sperm AF and 95% CIs (estimated fraction ± binomial CI, based on fraction of mutant reads; see Supplementary Dataset 7) for all mosaic variants. Dotted lines demark recurrent variants that suggest parental mosaicism. **h**, Number of mosaic dSTRΔs found in each father. **i**, Exemplary dSTRΔ in F05, where the child had an expansion of a tetranucleotide repeat (TCTA) on the paternal haplotype (12–13×) based on bulk sequencing. **j**, TCTA repeat AFs from 200× WGS data for paternal blood and sperm demonstrated PGM for the 13× variant.

enriched/BDE) and BDO (Fig. 4b). We identified 62 SDO, 61 BSS and 568 BDO, the last of these probably reflecting clonal hematopoiesis[23] primarily arising from the father of F02 (Fig. 4c). There were 9–23 variants in the sperm of each father, all with the potential to transmit to an offspring.

The AF of PGM variants ranged from a maximum of ~35% to the lower limit of detection, ~1.5% (Fig. 4d). Compared with sperm AF, blood AF showed two trends: at higher sperm AFs, blood AFs were similar to sperm AF while at lower sperm AFs, blood AFs were very low or undetectable (Extended Data Fig. 9d–f).

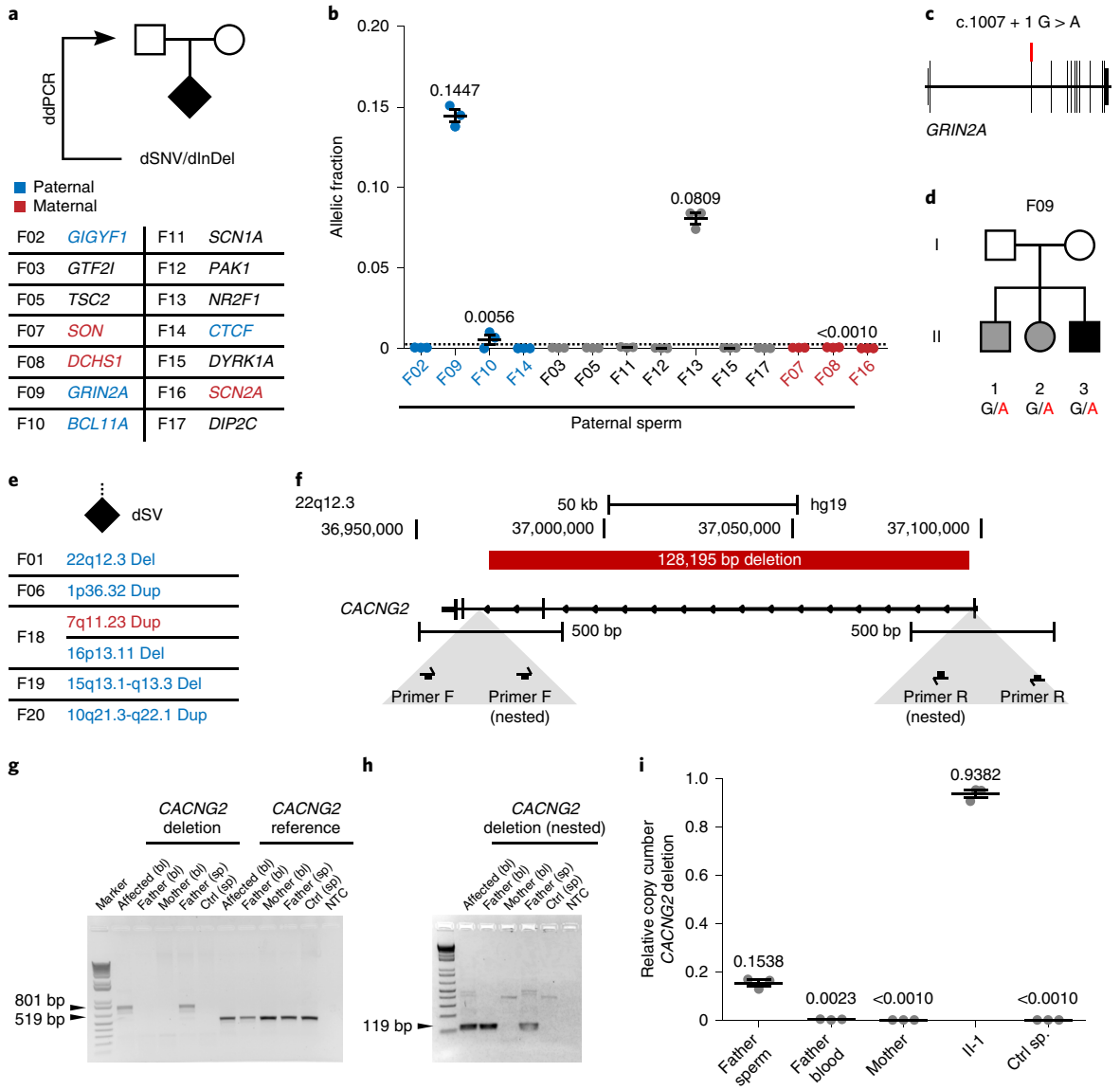


Fig. 3 | Pathogenic ASD DNMs benefit from risk stratification through sperm sequencing. a, Fourteen ASD families with a causative dSNV/de novo insertion/deletion (dInDel) in the child and phased, where possible, to the parental haplotype (blue, paternal; red, maternal). Gonadal mosaicism was assessed by ddPCR for each dSNV/dInDel. **b**, AF (determined by ddPCR) of mutant alleles in paternal sperm ($n = 3$ experimental replicates for each sample; shown are mean ± s.e.m. and individual values). **c**, Schematic of *GRIN2A* and the PGM variant found in F09. **d**, Pedigree of family F09. Black, ASD; gray, epilepsy with ADHD symptoms. All three children shared the *GRIN2A* G>A mutation. **e**, Five ASD families with a causative dSV. Haplotype was determined from the WGS data as either paternal (blue) or maternal (red). Only the 22q12.3 deletion in F01 showed gonadal mosaicism, as also described in Fig. 2c,d. **f**, Genomic *CACNG2* locus (22q12.3) and the pathogenic 128,195-bp deletion in F01. Below: primers for nested PCR used for deletion detection. **g**, Agarose gel for the primary PCR products from blood (bl) and sperm (sp). *CACNG2* deletion: 801-bp band detected in DNA from affected and paternal sperm; 519-bp reference band detected in all samples as a control. **h**, Agarose gel for nested PCR products (arranged as in **g**). **g**,**h**, Representative gels from two independent replicates. **i**, ddPCR showed CN mosaicism at 0.1538 copies or ~7.5% AF in sperm, 0.0023 copies or ~0.1% in blood from father, 0.9382 copies or ~47.0% AF in blood from the affected individual, and undetectable in samples from mother and control ($n = 3$ experimental replicates for each sample; shown are mean ± s.e.m. and individual values).

This suggested two separate origins of PGM during paternal embryogenesis—the former occurring before and the latter after germ cell specification. The AF distribution of SDO, BSS and BDO was consistent with this model, where most SDO variants occurred at AFs <10%, whereas BSSs showed an AF range up to 35% (Fig. 4e,f). BDO AFs tended to mimic those of SDO, but there was a distribution tail with higher AFs probably reflecting clonal hematopoiesis. These BDO variants, while numerous, had little chance of being transmitted to an offspring because they were absent in sperm. Therefore, sequencing of blood only to identify potentially transmissible variants would not distinguish BDO from BSS and would miss SDO variants completely.

Mutational signatures suggest an embryonic origin of PGM. We then combined all mosaic SNVs detected in both approaches (Figs. 1 and 4) to observe common patterns for these variants. While there was no clustering along, or enrichment across, chromosomes (Extended Data Fig. 10a,b and Supplementary Dataset 9), we observed distinct mutational signatures differentiating variant classes. Assessing the relative contribution of each of the six possible

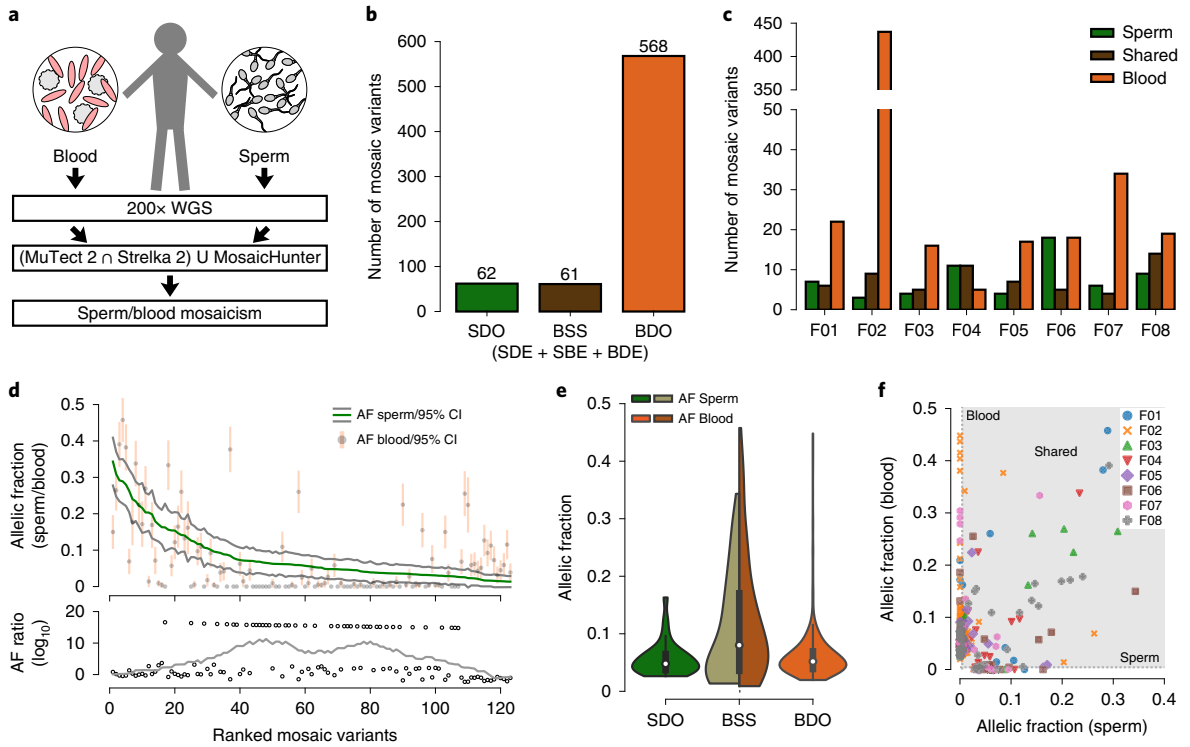


Fig. 4 | Sperm sequencing reclassifies risk for ~50% of transmittable mosaic variants. a, Blood and sperm from eight fathers were subjected to 200× WGS followed by detection of mosaicism using the intersection of MuTect 2 and Strelka 2 and union with MosaicHunter. **b**,**c**, Total number of mosaic variants (**b**) and those found in each father (**c**) that were SDO, BSS or BDO. F02 showed a substantially increased number of BDO variants, most probably related to clonal hematopoiesis/collapse due to his advanced age at sampling (70 years). **d**, Ranked plot of estimated sperm and blood AF with 95% CIs (estimated fraction ± binomial CI, based on fraction of mutant reads; see Supplementary Dataset 8) for all 123 gonadal mosaic variants detected as mosaic in sperm. Lower plot shows the $\log_{10}$ transformed ratio of sperm and blood AFs (0 replaced by 1×10^{-8}) and the rolling average over 20 data points to display the local trend. **e**, Violin plots with inner box plots (showing median and quartiles) of AFs of all three types of variant as indicated ($n=62$ SDO variants, 61 BSS and 568 BDO). **f**, Sperm versus blood AF for all detected mosaic variants, coded by individual. BDO and BSS mosaic variants reached higher AF than SDO. Gray area denotes region of variants detectable in both sperm and blood.

base substitutions, mosaic variants differed from the background of Genome Aggregation Database (gnomAD) variants in several categories (Fig. 5a,b and Extended Data Fig. 10c). The early shared BSS mosaics differed from SDO and BDO variants, which were similar to each other. Supporting an embryonic origin of these variants, they were all depleted in T>C variants, a class that was correlated with environmental damage and aging gonads and depleted in variants that were shared among siblings[6,10]. The differential signals for BSS variants enriched in C>A and T>G mutations relative to gnomAD and SDO and BDO mosaics are consistent with distinct mutational mechanisms in early embryonic development compared to those at later stages[14,24].

Discussion

Our results represent a significant improvement over previous strategies, where assessment of parental blood mosaicism only was used in combination with population statistics[7,10]. The role of sperm mosaicism has been increasingly recognized in single-gene disorders[25–28], and our work complements these efforts by providing a more general assessment of sperm mosaicism. Our data suggest a model of three major types of PGM (Fig. 4c): type I arises during the terminal differentiation of sperm and never recurs. Type II arises in proliferating spermatogonial stem cells (SSCs) and includes those that are extant clonally (IIa) or those under positive selection (IIb), akin to the 'selfish sperm' hypothesis[29]. Type IIa probaby represents mutations accumulating in individual SSCs and proposed to underlie the increased mutational load with age[10,11], although its importance in this process is controversial[24]. Multiple inheritance is rare for IIa whereas IIb is similar to IIa because they have the same origin, but their selective advantage results in overproliferation of the SSC clone and the potential for population-wide recurrence.

Type III arises during paternal embryonic development, before primordial germ cell (PGC) specification or within the PGC population, and may be detectably mosaic in sperm, resulting in the potential for recurrence. The timing of a mutation probably determines its abundance and patterns of mosaicism between sperm and somatic tissue, and our data suggest distinct mutational mechanisms between BSS and SDO variants.

Employing our methods, a distinction between the contribution of type I and type II mosaicism to the male-specific mutational burden is not possible, as both are below the detection limits; similarly, type III PGM occurring after PGC specification is probably not possible unless it is positively selected[27]. In contrast, our work focuses on the detection of type III mosaicism, which can stratify the risk of recurrence. Considering the fraction of mosaic variants detected for each father, we estimate that on average 2.9% (95% CI: 1.4–4.4%) of variants fall within this category, and this increases to 4.3% (95% CI: 1.6–7.1%) if a variant can be phased to the paternal haplotype. Nevertheless, based on our data and those of previous, gene-centric studies on sperm mosaicism, even within this group, risk can vary by an order of magnitude[25,27,28].

Thus, the patterns of sperm mosaicism and the resulting framework for its detection that we present have the potential to impact clinical testing in two ways. First, direct assessment of previously

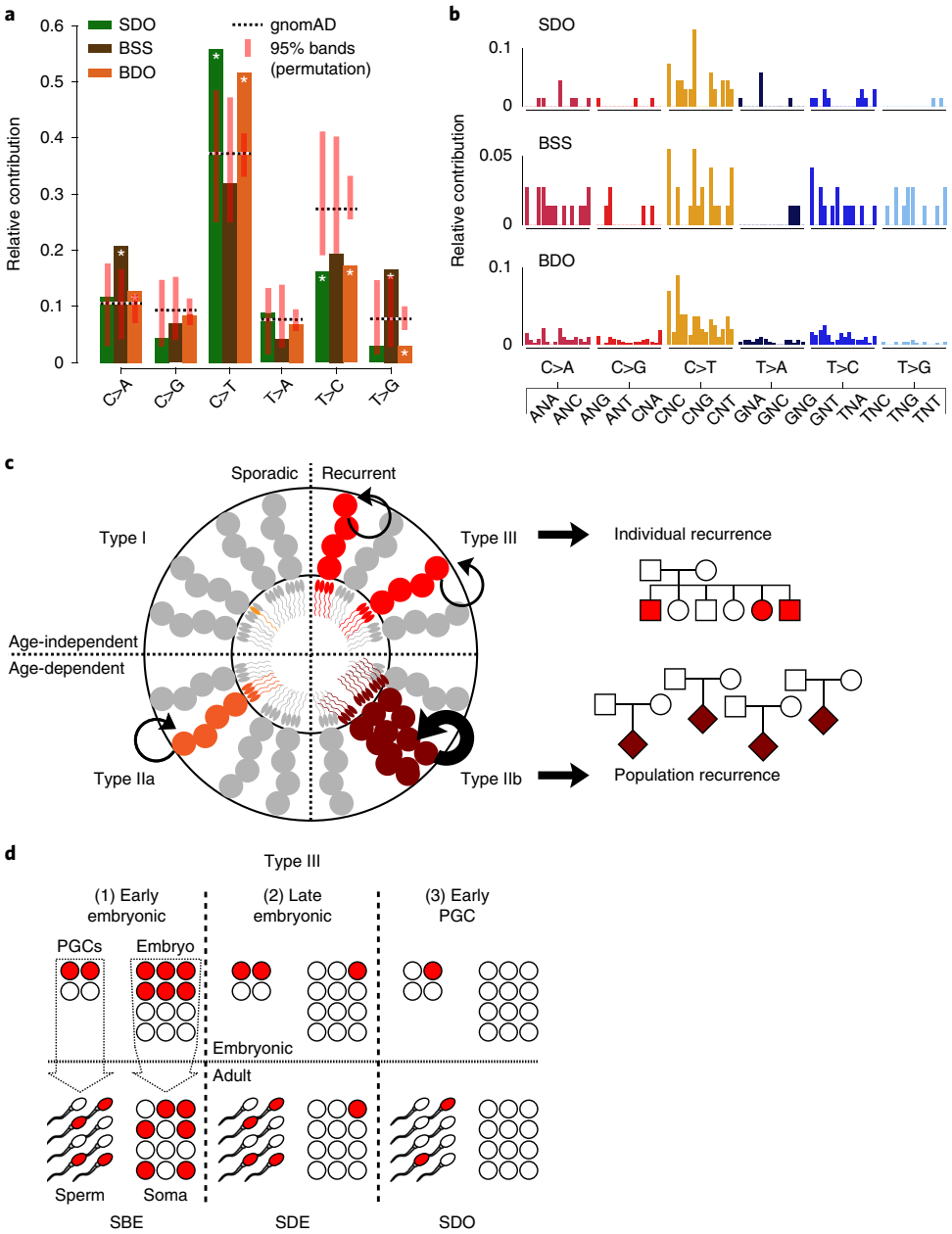


Fig. 5 | Sperm mosaic variant mutation patterns support a developmental origin. a, Mutational signatures (six categories) for the three classes of mosaicism compared to the overall gnomAD signature and a permuted subset (n = 1,000 permutations for n = 68 (SDO), 72 (BSS) and 568 (BDO) gnomAD SNVs; shown is the 95% band). Asterisks indicate observed signatures that lie outside the 95% band of the permuted variants. SDO and BDO showed signatures that differed from gnomAD and the BSS variants; BSS variants likewise showed a mutational signature that was distinct from the gnomAD population. **b**, Mutational signatures (96 categories; trinucleotide environment) of the three classes of mosaicism. **c**, Model for four types of PGM from testis tubule cross-section, with SSCs at the perimeter and mature sperm in the lumen. Types I and IIa PGM occur in a single sperm (I) or SSC (IIa), thus contributing to only a fraction of total sperm and associated with nonrecurrent disease. Types IIb and III mutations lead to selective growth advantage and elevate population-level recurrence risk (IIb), or occur during paternal embryogenesis, leading to multiple independent mutant SSCs and association with mutational recurrence (III). **d**, Type III PGM mosaicism occurs (1) during early paternal embryogenesis, seeding sperm and somal progenitors at equally high AFs; (2) during late embryogenesis, seeding stochastically at variable AFs between tissues; or (3) in early PGC differentiation, seeding only gonads. Note: PGCs are the early embryonic progenitors of SSCs.

transmitted pathogenic variants in paternal sperm allows for the stratification of fathers with low and high recurrence risk through TAS or ddPCR analysis. Second, even without any previous risk or family history, prospective fathers who may want to know their risk of transmitting a high-impact variant to their child could undergo deep sequencing of their sperm, followed by mosaic analysis of these data. This potential is highlighted by our finding that one of the SDO variants (F06: chr9:g.131380333 G>A; NP_001123910.1:p.Arg1849Gln; 3.7% AF) was located in *SPTAN1*, a gene known to cause infantile epileptic encephalopathy (MIM: 613477)[30]. While this specific variant has not previously been reported, it was predicted to be 'potentially disease-causing' by MutationTaster[31], had a MutPred2 score of 0.687 (ref. [32]) and a different nonsynonymous change in this same amino acid residue has been reported in

affected children in ClinVar (SCV000243194.10, SCV000553140.2; p.Arg1849Trp). Based on our results, we would predict that this variant, which has the potential to be pathogenic, has a 3.7% inheritance risk for any subsequent child of the father in F06.

There are still several limitations and impediments regarding the application of sperm mosaicism testing. First, both approaches require the assessment of suspected high-penetrance variants and currently ignore modifiers and polygenic risk scores. This limitation is exemplified by the *GRIN2A* variant in family F09, where it is unclear whether the variability in expressivity is due to environment, genetic modifiers or stochasticity[15,16,33]. Second, the absence of detectable mosaicism in paternal sperm can stratify the family into low risk only if the mutation of interest has been phased to the paternal haplotype. While phasing can be achieved through several experimental approaches[34,35], including the nanopore sequencing we present in this manuscript, its implementation in clinical practice is still uncommon. Third, while we show examples of resampling for three of the pathogenic variants and the relative stability of mosaicism between samples, it is unclear whether this is true across all mosaic variants and should be studied systematically in future. However, it is a problem that would be less relevant when testing sperm samples that are directly used for in vitro fertilization. Finally, our framework for the unbiased detection of mosaicism is tuned for specificity and may therefore miss clinically relevant variants. Similarly to mosaic analysis of cancer, implementation of sperm analysis for mosaic risk mutation has to be tuned for clinical application and may require large-scale secondary validation by methods such as TAS.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-019-0711-0.

Received: 15 August 2019; Accepted: 21 November 2019;
Published online: 23 December 2019

References

1. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
2. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 e712 (2017).
3. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
4. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
5. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
6. Jonsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
7. Campbell, I. M. et al. Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. *Am. J. Hum. Genet.* **95**, 345–359 (2014).
8. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17**, 241 (2016).
9. Freed, D., Stevens, E. L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes (Basel)* 5, 1064–1094 (2014).
10. Jonsson, H. et al. Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
11. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
12. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
13. Brandler, W. M. et al. Frequency and complexity of de novo structural mutation in autism. *Am. J. Hum. Genet.* **98**, 667–679 (2016).
14. Huang, A. Y. et al. Distinctive types of postzygotic single-nucleotide mosaicisms in healthy individuals revealed by genome-wide profiling of multiple organs. *PLoS Genet.* **14**, e1007395 (2018).
15. Carvill, G. L. et al. GRIN2A mutations cause epilepsy-aphasia spectrum disorders. *Nat. Genet.* **45**, 1073–1076 (2013).
16. Lemke, J. R. et al. Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. *Nat. Genet.* **45**, 1067–1072 (2013).
17. Turner, D. J. et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40**, 90–95 (2008).
18. Hehir-Kwa, J. Y. et al. De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48**, 776–778 (2011).
19. Escaramis, G., Docampo, E. & Rabionet, R. A decade of structural variants: description, history and methods to detect structural variation. *Brief. Funct. Genomics* **14**, 305–314 (2015).
20. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
21. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
22. Huang, A. Y. et al. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res.* **45**, e76 (2017).
23. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
24. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).
25. Bernkopf, M. et al. Quantification of transmission risk in a male patient with a FLNB mosaic mutation causing Larsen syndrome: implications for genetic counseling in postzygotic mosaicism cases. *Hum. Mutat.* **38**, 1360–1364 (2017).
26. Hancarova, M. et al. Parental gonadal but not somatic mosaicism leading to de novo NFIX variants shared by two brothers with Malan syndrome. *Am. J. Med. Genet. A* **179**, 2119–2123 (2019).
27. Wilbe, M. et al. A novel approach using long-read sequencing and ddPCR to investigate gonadal mosaicism and estimate recurrence risk in two families with developmental disorders. *Prenat. Diagn.* **37**, 1146–1154 (2017).
28. Yang, X. et al. Genomic mosaicism in paternal sperm and multiple parental tissues in a Dravet syndrome cohort. *Sci. Rep.* **7**, 15677 (2017).
29. Goriely, A. & Wilkie, A. O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
30. Hamdan, F. F. et al. Identification of a novel in-frame de novo mutation in SPTAN1 in intellectual disability and pontocerebellar atrophy. *Eur. J. Hum. Genet.* **20**, 796–800 (2012).
31. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
32. Pejaver, V. et al. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. Preprint at *bioRxiv* https://doi.org/10.1101/134981 (2017).
33. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
34. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
35. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).

Methods

Binomial modeling of detection threshold. Depicted curves were based on a classic binomial model assuming that the AF of a mutation represents the probability of encountering a mutant read. The cumulative probability was calculated using the integrate.quad function of the scipy module from Python.

Simulation and analysis. To determine our sensitivity to detect mosaic variants, we created simulated datasets that contained known mosaic variants at low frequencies. We first randomly generated 10,000 variants from chromosome 22 as our set of mosaic variants. We then used Pysim[36] to simulate Illumina paired-end sequencing reads from reference chromosome 22 and a version of chromosome 22 that contained the alternate alleles from our 10,000 mosaic variants. These two sets of reads were then combined to create a series of datasets with mosaic variants at 1, 2, 3, 4, 5, 10, 15, 20, 25 and 50% AF. The coverage of these datasets was 200×. We processed these reads through our standard mapping and somatic variant-calling pipelines (see below), and calculated sensitivity to detection of mosaic variants at each AF as the fraction of simulated variants called by our dSNV pipeline, or by both MuTect 2/Strelka 2 and MosaicHunter.

Patient recruitment. Patients were enrolled, according to approved human subjects protocols at the University of California, for blood, saliva and semen sampling. Semen was collected for all fathers of families F01–20. For F09–12, saliva from the fathers and their family members was obtained; for F01–08 and F13–20, DNA from blood was extracted. WES trio analysis for F09–12 was performed on DNA extracted from lymphocyte cell lines (generated by the NIMH Repository) and results were confirmed in saliva samples. WGS trio analysis for F01–08 and F13–20 was performed on DNA derived from blood. Each father provided a single sperm sample, with the exception of F01, F09 and F13, where a second sample was obtained 1, 3.5 and 4 months, respectively, after the first. Patients were part of two independent cohorts, assembled to identify dSNVs and dSVs through trio sequencing[1,4]: the REACH cohort[12,13], consisting of 265 families with a proband with general features of ASD and recruited at Rady Children's Hospital San Diego (J.S.) and at Mutua Terrassa Hospital Barcelona (M.A., A.H. and J.S.), and one focusing on 98 probands with ASD and an additional diagnosis of epilepsy, recruited at NYU Medical School (O.D. and J.G.G; unpublished). The REACH cohort has been described previously[12,13]. The cohort assembled by J.G.G. and O.D. represents a new recruitment effort that focused on patients with a diagnosis of ASD with associated epilepsy. Patients were evaluated by a child neurologist and a clinical geneticist for general and neurological assessment after referral from their primary care physician for concern about developmental delay and autism. Intellectual function was assessed by IQ score. Speech was assessed by a speech therapist fluent in the child's native language. Brief videos of each affected member were collected during the examination as part of the clinical assessment. Autism was assessed by a clinical psychologist using the Autism Diagnostic Interview–Revised, the Autism Diagnostic Observation Schedule and the Childhood Autism Rating Scale, and developmental milestones were assessed with the Vineland Scale and hyperactivity with the Conners Parent/Teacher Scale, all administered in the child's native language by a trained psychologist. Epilepsy was assessed by a trained specialist and included history of daytime and night-time seizures, seizure types, length, onset and resolutions, and treatment history. Electroencephalography (EEG) was assessed awake and asleep using a minimum of 21 electrodes and '10 to 20' system placements as recommended by the International Federation of Clinical Neurophysiology. In specific subjects, a 24-h EEG was recorded to evaluate the possibility of night-time seizures and to identify seizure foci. All subjects were recruited between the ages of 3–8 years. Patients were followed longitudinally to assess response to anticonvulsant therapy and behavioral therapy. All patients were seen at the NYU School of Medicine and were recruited through the ethical framework at the University of California, San Diego.

Blood and saliva extraction. DNA was extracted on an Autopure LS instrument (Qiagen).

WES and WGS trio analysis. WGS sequencing and analysis for F01–08 and F13–20 were performed as described previously[13,37]. Exome capture and sequencing of F09–12 were performed at the New York Genome Center (Agilent Human All Exon 50 Mb kit, Illumina HiSeq 2000, paired-end, 2 × 100) and the Broad Institute (Agilent Sure-Select Human All Exon v.2.0, 44-Mb baited target, Illumina HiSeq 2000, paired-end, 2 × 76). Sequencing reads were aligned to the hg19 reference genome using BWA (v.0.7.8). Duplicates were marked using Picard's MarkDuplicates (v.1.83, http://broadinstitute.github.io/picard) and reads were realigned around insertion/deletions (InDels) with GATK's IndelRealigner. Variant calling for SNVs and InDels was performed according to GATK's best practices by first calling variants in each sample with HaplotypeCaller and then jointly genotyping them across the entire cohort using CombineGVCFs and GenotypeGVCFs. Variants were annotated with SnpEff (v.4.2) and SnpSift (v.4.2), and allele frequencies from the 1000 Genomes Project and the Exome Aggregation Consortium (ExAC)[38]. De novo variants were called for probands using Triodenovo (v.0.06) with a minimum de novo quality score of 2.0 and subjected to manual inspection. Variants from F01–F08 were further interrogated for postzygotic mosaic variants (PMVs) that might be present in the children[9,39]. Among all 912 variants, only four showed significant deviation from an expected 0.5 AF using a binomial model; this effect was seen before multiple testing and disappeared following Bonferroni correction. This lack of PMVs in our data is most probably a reflection of limited sequencing depth (~40×) and cannot conclusively exclude the existence of PMVs in our data. Nevertheless, conservatively, we assumed that all 912 dSNVs were true DNMs. We further interrogated F01–08 for possible paternally mosaic variants that might have been erroneously reported as inherited heterozygous variants; such artifacts might have resulted in an underestimation of mosaicism and overestimation of SDO and SDE variants. However, multiple filtering approaches did not result in the identification of any such variants. While we cannot exclude their existence, we believe that their contribution to mosaicism—if any—is minor in our dataset.

Sperm extraction. Extraction of sperm cell DNA from fresh ejaculates was performed as previously described[40]. In short, sperm cells were isolated by centrifugation of the fresh (up to 2-d) ejaculate over an isotonic solution (90%) (Sage/Origio, no. ART-2100; Sage/Origio, no. ART-1006) using up to 2 ml of the sample. Following a washing step, quantity and quality were assessed using a cell-counting chamber (Sigma-Aldrich, no. BR717805-1EA). Cells were pelleted and lysis was performed by the addition of RLT lysis buffer (Qiagen, no. 79216), Bond-Breaker TCEP solution (Pierce, no. 77720) and 0.2-mm stainless steel beads (Next Advance, no. SSB02) on a Disruptor Genie (Scientific Industries, no. SI-238I). The lysate was processed using reagents and columns from an AllPrep DNA/RNA Mini Kit (Qiagen, no. 80204). Concentration of the final eluate was assessed employing standard methods. Concentrations ranged from ~0.5–300 ng μl^{-1}.

WGS of matched sperm and blood samples. WGS was performed using an Illumina TrueSeq PCR-free kit (350-bp insertion) or a TrueSeq Nano kit (350-bp insertion) on an Illumina HiSeqX. Paired-end FASTQ files of deeply (~200×) sequenced blood and sperm samples from fathers were aligned to the hg19 reference genome (1000 Genomes v.37) with BWA mem (v.0.7.15-r1140), specifying the –M option that tags chimeric reads as secondary and that are required for certain downstream applications that implement this legacy option. The resulting average mean coverage was 227× for blood samples and 222× for sperm samples, with an average read length of 150 bp for both sets. Duplicates were removed with the markdup command from sambamba (v.0.6.6), and base quality scores were recalibrated with the Genome Analysis ToolKit (GATK v.3.5-0-g36282e4). SNPs and InDels were called with HaplotypeCaller jointly genotyping within pedigrees, consisting of the deep-coverage (~200×) genomes from the father's blood and sperm and ~40× coverage genomes derived from the blood of both parents and children.

Oxford Nanopore (ONP) sequencing and analysis. Whole-genome sequencing libraries were generated with ONP 1D-long reads for all children (except for F03-II-2, due to lack of sufficient DNA) in deep-whole-genome families (F01–F08) and a subset of families with pathogenic variants (F13–F15), according to the manufacturer's recommendations. FASTQs were aligned to the hg19 reference genome with BWA mem with the -x ont2d option for ONP reads. Coverage of proband samples ranged from 3× to 15× (average, 8.6×), with an average read length of 5,349 bp.

Haplotype phasing. To phase dSNVs, a set of phase-informative single-nucleotide polymorphisms (SNPs) from the WGS germline variant calls, or from an assembly of the local area using Nextera sequencing (see below) of a 20-kb region around the dSNV, was determined. Phase-informative SNPs were those where the child was heterozygous and either (1) one parent was heterozygous or homozygous for the alternate allele while the other was homozygous for the reference allele, or (2) one parent was heterozygous while the other was homozygous for the alternate allele. Second, where applicable, long reads (ONP reads, average length 5,349 bp) were identified that contained both a dSNV and one or more phase-informative SNPs. The number of dSNV and phase-informative SNP combinations that were present in reads and consistent with the dSNV occurring on a maternal or paternal haplotype were counted. Reads containing an InDel flanking either the dSNV or the phase-informative SNP were excluded from the analysis. Finally, dSNVs were assigned to maternal and paternal haplotypes if there were: (1) a minimum of two counts and (2) the haplotype with the majority of counts had at least two-thirds of total counts. For F09–F12, F16 and F17, we attempted phasing using a Drop-Phase approach[41]. In short, a complementary assay to the mutant allele at the dSNV position was designed for both the wild-type and the variant allele (see ddPCR design, validation, and setup of experiments for SNV analysis). Co-occurrence of the mutant dSNV was then ssessed for both genotypes and quantified as described previously[41] (Supplementary Dataset 8).

Sanger sequencing of SNVs. PCR and Sanger sequencing were performed according to standard methods. Primer sequences can be found in Supplementary Dataset 10. Validated mutations and surrounding SNPs were also used as basis for the design of ddPCR assays, where applicable.

ddPCR design, validation and setup of experiments for SNV analysis. Using the Primer3Plus web interface[42–44], the amplicon and probes for wild-type and mutant were designed to distinguish reference and alternate allele (settings are given in Supplementary Information under Additional information). Probes were required to be located within 15 bp up- and 15 bp downstream of the mutation and adjusted, so melting temperatures were matched between reference and alternate probe. In addition, if possible, amplicons were kept at 100 bp or shorter and probes at 20 bp or shorter. Specificity of the primers was assessed using Primer-BLAST. Custom primer and probe mixes (primer/probe ratio of 3.6) were ordered from IDT with FAM-labeled probes for the alternate, and HEX-labeled probes for the reference, allele (Supplementary Dataset 10). Optimal annealing temperature, specificity and efficiency were tested using custom gblocks (IDT) or patient DNA at a range of dilutions. ddPCR was performed on a Bio-Rad platform using a QX200 droplet generator, a C1000 touch cycler, a PX1 PCR Plate Sealer and a QX200 droplet reader, with the following reagents: ddPCR Supermix (Bio-Rad, no. 1863024), droplet generation oil (Bio-Rad, no. 1863005), cartridge (Bio-Rad, no. 1864008) and PCR plates (Eppendorf, no. 951020346). Aiming for 30–60 ng per reaction, up to 8 µl of DNA solution was used in a single reaction. Data analysis was performed using the software packages QuantaSoft and QuantaSoft Analysis Pro (Bio-Rad). Each run included technical duplicates or triplicates (as indicated in figure legends). For direct comparison of sperm samples we used seven technical replicates, except for F09 where the total amount of sperm DNA was limiting. Across all ddPCR reactions that were designed for SNV detection, we determined that the minimum AF that could reliably be detected was 0.1%. Therefore, we set this as the threshold of detection. Raw data for ddPCR experiments can be found in Supplementary Dataset 3.

TAS. PCR products for sequencing were designed with a target length of 160–190 bp, with primers being at least 60 bp distant from the base of interest. Primers were designed using the command-line tool of Primer3 with a Python wrapper (Supplementary Dataset 10). PCR was performed according to standard procedures using GoTaq Colorless Master Mix (Promega, no. M7832) on sperm, blood and an unrelated control. Amplicons were either enzymatically cleaned with ExoI (NEB, no. M0293S) and SAP (NEB, no. M0371S) treatment or gel extraction (Zymo Research, no. D4007) where necessary. Following normalization with the Qubit HS Kit (Thermo Fisher Scientific, no. Q33231), amplification products were processed according to the manufacturer's protocol with SureSelect SPRI Beads (Beckman Coulter, no. A63881) at 1.2×. Library preparation was performed according to the manufacturer's protocol using a Kapa Hyper Prep Kit (Kapa Biosystems, no. KK8501) and barcoded independently with unique dual indices (IDT for Illumina, no. 20022370). After sequencing on an Illumina HiSeq 4000 with 100-bp paired-end reads, reads were mapped to the hg19 reference genome (1000 Genomes v.37) and processed according to GATK v.3.8 best practices. Across all amplicons, read numbers (mean ± s.d.) were 636,636 ± 382,226 in sperm, 831,556 ± 530,332 in blood and 857,289 ± 570,612 in control. Overall, read depth reached between 93× and 3,138,968×, with 99% >261× and 95% >1,809×. Putative mosaic sites were retrieved using samtools mpileup and pileup filtering scripts described in previous TAS pipelines[28,45]. Variants were considered validated if (1) their lower 95% CI boundary was above the upper 95% CI boundary of the control and (2) their AF was >0.5%.

Mosaic dSNV analysis. Using the read depth information generated by HaplotypeCaller, the AF for previously called dSNVs was determined. Additionally, dSNVs that fell within repetitive regions of the human genome were annotated using the repeatMasker (rmsk.txt) file from the University of California, Santa Cruz (UCSC). Variants that were homozygous alternate in the father and heterozygous in the proband, as well as those that were present in both blood and sperm at AFs that suggested an inherited heterozygous SNP (that is, AF >35% in both blood and sperm), were removed. Variants were further filtered to include only those with a gnomAD frequency <0.01 (ref. [46]). Mosaic variants were categorized based on their presence or absence in sperm and blood (≥3 reads minimum requirement in one of these; if ≥ three reads were present for one, the other only had to show ≥1 read). The three-read minimum was based on an expected Illumina per-base error rate of Q30/0.1% (that is, ~0.033% error rate to substitute to the expected dSNV). Given a read depth of 200×, a minimum of one read as evidence would result in a falsely assigned mosaic variant with ~6.5% probability, while with two reads this drops to ~0.2% and with three reads to ~0.005%. Given that the number of interrogated dSNVs is ~1,000, this would result in ~60, ~2 and ~0.05 false-positive variants, respectively. To be called sperm enriched, a variant's AF had to be three times higher in sperm than in blood ($\alpha > 3$), which was an arbitrarily determined threshold based mainly on the level of the 95% CI at ~200× at low AF. To assess the sensitivity of 200× WGS, in one family we also performed Multiplex Accurate Sensitive Quantitation (MASQ) on sperm DNA, which is capable of detecting AFs as low as 10^{-4}–10^{-6} (A.B.M. and Z.W., unpublished). We selected dSNVs where sperm WGS did not suggest gonadal mosaicism for the majority of variants, to assess whether additional mosaics might be identified. From 73 such dSNVs in F01, MASQ assay design was successful for 23, two of which were already detected as mosaic from 200× WGS of sperm. MASQ confirmed these two, but did not detect any additional variants that were mosaic in sperm. Only one of these variants remained unphased, confirming that the remaining 20 paternally derived dSNVs, even at this level of detection, were not found to be mosaic in sperm and that they probably arose either later in the sperm lineage, zygotically or postzygotically.

MuTect 2/Strelka 2 and MosaicHunter mosaic variant calling. Sperm- and blood-specific SNVs were called in the 200× WGS data using two somatic variant callers with default parameters, MuTect 2 (v.2.1)[20] and Strelka 2 (v.2.9.2)[21], setting the sperm sample as 'tumor' and the blood sample as 'normal' and vice versa. High-confidence calls for somatic mosaicism for each sperm–blood and blood–sperm comparison were performed by taking the intersection of variants identified by both callers (MS). These candidates were further filtered to reduce potential false positives as follows: we removed those that fell into repetitive regions, those that fell within 5 bp of a germline InDel, those that were part of a homopolymer or dinucletide repeat, and those that were present in gnomAD at allele frequencies >0.01. The latter filter was employed, because common variants that appear to be mosaic are most often artifacts. Shared mosaic variants (and some tissue-specific variants missed by MS) were called using the whole-genome single mode provided by MosaicHunter v.1.0 (ref. [22]) as previously described[14] (MH). Additionally, variants: (1) had to be within the fifth and 95th percentile ($76 <$ read depth < 280) for sequencing depth across all variants to control for artifacts; (2) had to be absent or at an allele frequency <0.01 in gnomAD; (3) could not be recurrent in our data set; (4) had to have a major allele consistent with the reference allele in hg19; and (5) had to have an AF <30% in at least one tissue (to remove probable heterozygous calls). If a mosaic variant was found only in one tissue by MH, that variant was determined as being a shared mosaic only if there were ≥3 reads supporting the alternative allele in the second tissue. Calls from both methods (MS and MH) were then combined to obtain tissue-specific and shared mosaicism.

Assessment of location of genome-wide distribution of mosaic variants. To assess the distribution of mosaic variants along the chromosomes, an equal number of variants (for mosaic dSNVs and unbiased calls that were sperm-specific, sperm mosaic, blood mosaic or blood-specific) was randomly generated with BEDTools from the called region from Strelka 2 with or without subtraction of the repeatMasker (rmsk.txt) file from UCSC as appropriate. This process was repeated 10,000 times to generate a distribution of the mean and s.d. of the distance of neighboring variants according to a broken-stick model[47].

Mutational signatures. Mutational signatures were determined for each variant by retrieving the trinucleotide sequence context using Python with pysam, and plotting the trans- or conversion based on the pyrimidine base of the original pair similar to previous studies[48]. gnomAD mutational signatures were obtained by retrieving SNVs present in the publicly available variant call format (VCF) file. To obtain a 95% band of expectation, an equivalent number of variants was randomly chosen from the gnomAD VCF. This process was performed for a total of 1,000 times to obtain a distribution and the 2.5th and 97.5th percentile of the simulated mutational signatures. Significance was reported if a mutational signature was outside the permuted 95% bands.

Mosaic dSV analysis of WGS data. We searched for evidence for mosaicism of structural variants in the fathers using depth of coverage, split-reads, discordant paired-ends and B-allele frequency in deeply sequenced paired-end genomes. Depth of coverage was estimated as the median per base-pair coverage within the SV locus, while omitting positions that overlapped assembly gaps, RepeatMasker elements, short tandem repeats and segmental duplications. We estimated copy number by dividing the median depth of coverage by the median coverage of the chromosome and multiplying by the ploidy number (2 for autosomes). Standard deviation of copy number was calculated by sampling the estimated copy number of 1,000 random nonoverlapping regions of the same length of the dSV, ensuring that each region did not overlap >50% to exclude elements listed above. Since the reported copy number of the de novo duplication appears to be elevated in related noncarriers, we opted to estimate the s.d. of copy number while controlling for repeat content. Hence, sampled regions contained a RepeatMasker content in the range 30–35% (dSV = 33%). Split-reads (also known as chimeric reads) are those with multiple alignments to the genome. Generally, if a read spanned a deletion or tandem duplication breakpoint, two alignments were generated with each segment mapping to opposite ends of the breakpoint. Similar to split-reads, discordant paired-ends had read fragments that span the SV breakpoint but the SV breakpoint resided in the unsequenced insert of the fragment. Consequently, the paired-ends mapped to opposite ends of the breakpoint producing an insert size approaching the size of the SV. We searched ±250 bp from the predicted breakpoint for SV-supporting reads, which were unique reads that were either split or contained discordant paired-ends with breakpoints that overlap at least 95% reciprocally to the SV. We reported the proportion of supporting reads to noninformative reads (those that do not support the SV) within the ±250-bp windows, which roughly estimates the proportion of mosaicism. Additionally for the de novo duplication SV, we searched for deviations in B-allele frequency defined as the proportion of reads that support the alternate variant to all reads covering the variant in question.

Normalized sequencing depth calculations generated by CNView[49] were derived from binned coverages in 45-kb, nonoverlapping windows.

dSTRΔ calling and mosaicism detection. Analysis of STR expansions and contractions were performed using HipSTR[50] (v.0.6) jointly on all BAM files (40× trios and >200× blood and sperm of fathers). The reference STR set provided by HipSTR for GRCh37 (GRCh37.hipstr_reference.bed) and default options were used, except for -def-stutter-model and -output-gls. Furthermore, a modified version of HipSTR's denovofinder tool was run on each of the 40× trios. The posterior probability of a de novo mutation was calculated using HipSTR genotype likelihood and STR loci mutation rates as priors. Strict quality filters to detect de novo STRs were applied within trios. STR loci were excluded from analysis if they were in segmentally duplicated (UCSC hg19.genomicSuperDups table)[51,52] regions. Genotype STR calls in all family members were required to have a minimum genotype quality of 0.9, a maximum of 15% of reads with stutter or InDel, at least ten spanning reads and at least 20% of reads to support each allele. STR loci were excluded if homozygous in the child or if they contained homopolymers and dinucleotide repeat motifs. De novo STR mutations were further required to have a posterior probability of de novo mutation ≥0.8. Mutations were excluded if they were not a multiple of the repeat motif unit, or if the de novo allele was found in one of the parents at >0.1 allele frequency. STR mutations were further considered only if the repeat unit was ≥3, because homopolymers and dinucleotide repeats were enriched for false-positive calls. The remaining loci were annotated with their phase where possible, and de novo allele frequencies in the >200× sperm and blood samples. dSTRΔs were qualified as inconclusive if mosaicism was detected in both mother and father, as true de novo if no mosaicism was detected in the parents, as maternal if mosaicism was detected in the mother only, and as paternal if mosaicism was detected in blood, sperm or both. In regard to dSNVs, sperm-enriched variants were annotated as such if the AF was >3-fold higher in sperm than blood. Phase of STR was inferred from genotype: if a unique allele was inherited from one of the parents, the STRΔ was assumed to be derived from the other.

Nextera sequencing to identify informative SNPs. PCR products for sequencing were designed to encompass 1 kbp for the assembly of the local region around the mutation for phasing of F09–12 (Supplementary Dataset 10). Parallelized primer design was achieved using the web interface of PCRTiler[53]. PCR was performed according to standard procedures using GoTaq Colorless Master Mix (Promega, no. M7832). Successful amplification products were processed according to the manufacturer's protocol with SureSelect SPRI Beads (Beckman Coulter, no. A63881) at 0.8–1.0×, a Nextera DNA Library Preparation Kit (Illumina, no. C-121-1031) and a Nextera Index Kit (Illumina, no. FC-121-1011). After sequencing on an Illumina MiSeq, reads were mapped and processed according to GATK best practices. Variants were called using GATK's HaplotypeCaller.

Mosaic SV analysis using PCR and ddPCR. Nested PCR was performed using blood DNA extracted from the F01 trio (proband, mother and father), as well as sperm from the F01 father and a nonrelated male. Primers were designed using Primer3Plus online software to span the deletion breakpoints within *CACNG2* as determined by WGS analysis within 500-bp windows up- and downstream of the predicted deletion. Additionally, a reverse primer was designed to be used with the nested forward primer as an amplification control (Supplementary Dataset 10). All PCR reactions were in 25-µl volumes and included 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2 mM $MgCl_2$, 1 U of Taq (Thermo Fisher Scientific) and 300 nM of each appropriate primer. DNA template was 50 ng of DNA from blood or sperm for the initial PCR (using the external set of primers), or 1 µl of the initial PCR product for the nested (internal) PCR. PCR reactions were run following a standard ramp speed protocol using a C1000 Touch Thermal Cycler (Bio-Rad), with cycling consisting of 2 min initiation at 95 °C, 35 cycles of 95 °C for 30 s, 55 °C anneal for 30 s and 72 °C for 1 min, followed by a final extension at 72 °C for 3 min. Products were resolved on 2% agarose gels. For ddPCR analysis, primer and probe sets for the SVs (copy number and break point analysis) were designed using Primer3Plus (Supplementary Text and Supplementary Dataset 10). Primers were designed to span the deletion breakpoints within the region or to lie within an intron of a centrally located gene within the deleted or duplicated region. Custom primer and FAM-labeled probe mix at a primer:probe ratio of 750/250 nM was ordered from either Bio-Rad or IDT as described above, as well as a HEX-labeled pre-validated copy number variation assay specific for *RPP30* as an internal control (assay ID: dHsaCP2500350). ddPCR was performed and analyzed as described above. Raw data for ddPCR experiments can be found in Supplementary Dataset 3.

Data processing. Data analysis and plotting were performed using GraphPad Prism, R and Python (pandas, matplotlib and seaborn modules).

Statistics. Statistical analyses and fitting were performed using GraphPad Prism or Python with the SciPy, Astropy or StatsModels modules, or calculated directly using Pandas for percentiles. Regression analysis was performed using a simple ordinary least squares model with StatsModels. We calculated 95% CIs around the estimated fraction (allelic fraction) as binomial confidence intervals based on the estimated fraction and read number for each data point. Mean and s.e.m. were calculated using GraphPad Prism's integrated function. GraphPad Prism was also used to perform the unpaired, two-tailed *t*-test and the two-tailed Mann–Whitney test. Permutation analyses are described above in the respective Methods sections. To calculate the mean population fraction of mosaic variants from the dSNV data, the mean and 95% CI were calculated from the eight fathers, for all variants and those that were paternally phased in each individual; 95% CIs were calculated using the statsmodels.stats.api.DescrStatsW() and the associated. tconfint_mean() functions.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Aligned BAM files generated for this study through deep WGS or TAS are available on SRA (accession no. PRJNA588332). WGS data used for de novo calling are available through the NIMH Data Archive (NDA; collection ID: 2019). Long-read sequencing data are likewise available on NDA (collection ID: 2795). NDA access is regulated by the standard organizational process and is subject to review by NDA. Data are also available through the corresponding authors upon reasonable request. Additionally, summary tables of the data are included as Supplementary Information.

Code availability

Algorithms used for mosaic variant detection were published previously. Any custom code is available through the corresponding authors upon reasonable request.

References

36. Xia, Y., Liu, Y., Deng, M. & Xi, R. Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinformatics* **18**, 53 (2017).
37. Michaelson, J. J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).
38. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
39. Krupp, D. R. et al. Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am. J. Hum. Genet.* **101**, 369–390 (2017).
40. Wu, H., de Gannes, M. K., Luchetti, G. & Pilsner, J. R. Rapid method for the isolation of mammalian sperm DNA. *Biotechniques* **58**, 293–300 (2015).
41. Regan, J. F. et al. A rapid molecular approach for chromosomal phasing. *PloS ONE* **10**, e0118270 (2015).
42. Untergasser, A. et al. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–W74 (2007).
43. Untergasser, A. et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
44. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
45. Xu, X. et al. Amplicon resequencing identified parental mosaicism for approximately 10% of 'de novo' SCN1A mutations in children with Dravet syndrome. *Hum. Mutat.* **36**, 861–872 (2015).
46. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at *bioRxiv* https://doi.org/10.1101/531210 (2019).
47. Goss, P. J. & Lewontin, R. C. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* **143**, 589–602 (1996).
48. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
49. Collins, R. L., Stone, M. R., Brand, H., Glessner, J. T. & Talkowski, M. E. CNView: a visualization and annotation tool for copy number variation from whole-genome sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/049536 (2016).
50. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
51. Bailey, J. A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
52. Karolchik, D. et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
53. Gervais, A. L., Marques, M. & Gaudreau, L. PCRTiler: automated design of tiled and specific PCR primer pairs. *Nucleic Acids Res.* **38**, W308–W312 (2010).

Acknowledgements

We thank the participants in this study for their contribution. M.W.B. was supported by an EMBO Long-Term Fellowship (no. ALTF 174-2015), which is co-funded by the Marie Curie Actions of the European Commission (nos. LTFCOFUND2013 and GA-2013-609409), and an Erwin Schrödinger Fellowship by the Austrian Science Fund (no. J4197-B30). This study was supported by grants to J.G.G. from the NIH

(nos. U01MH108898 and R01NS083823); the Simons Foundation Autism Research Initiative to J.G.G. (no. 571583), J.S. and M. Wigler (laboratory leader for A.B.M. and Z.W.); the NIH (nos. MH076431 and MH113715) to J.S.; and the Howard Hughes Medical Institute to J.G.G. Sequencing support was provided by the Rady Children's Institute for Genomic Medicine and ONP. O.D. acknowledges support from the Silverman Family Foundation and Finding a Cure for Epilepsy (FACES) and Seizures. We thank B. Hamilton, N. Chi, V. Stanley, A. Marsh, M. Wigler and L. Alexandrov for suggestions. We thank R. Sinkovits, A. Majumdar, S. Strande and the San Diego Supercomputer Center for hosting the computing infrastructure necessary for completing this project.

Author contributions

M.W.B., J.S. and J.G.G. conceived the project and designed the experiments. M.W.B., M.K., L.L.B., X.Y., S.A.W., C.A.B.G. and A.N. performed the experiments. D.A., R.D.G., I.M, X.Y., J.G., M.Gymrek, W.M.B., M.Gujral and M.W.B. performed the bioinformatic and data analyses. D.M., R.K. and E.S. performed de novo analysis of the cohort collected and provided by O.D. K.N.J., O.H., J.M.-V., M.C.C.B., J.U.F., M.C.P., M.A., A.H. and M.W.B. requested, organized and handled patient samples. A.B.M. and Z.W. performed the orthogonal sensitive detection of mosaic variants. M.W.B., J.G.G. and J.S. wrote the manuscript with input from R.D.G. and K.N.J. All authors saw and commented on the manuscript before submission.

Competing interests

M.W.B., D.A., M.K., K.N.J., W.M.B., J.S. and J.G.G. are inventors on a provisional patent (PCT ref. no. SD2017-181-2PCT) filed by UC, San Diego, titled 'Assessing risk of de novo mutations in males'.

Additional information

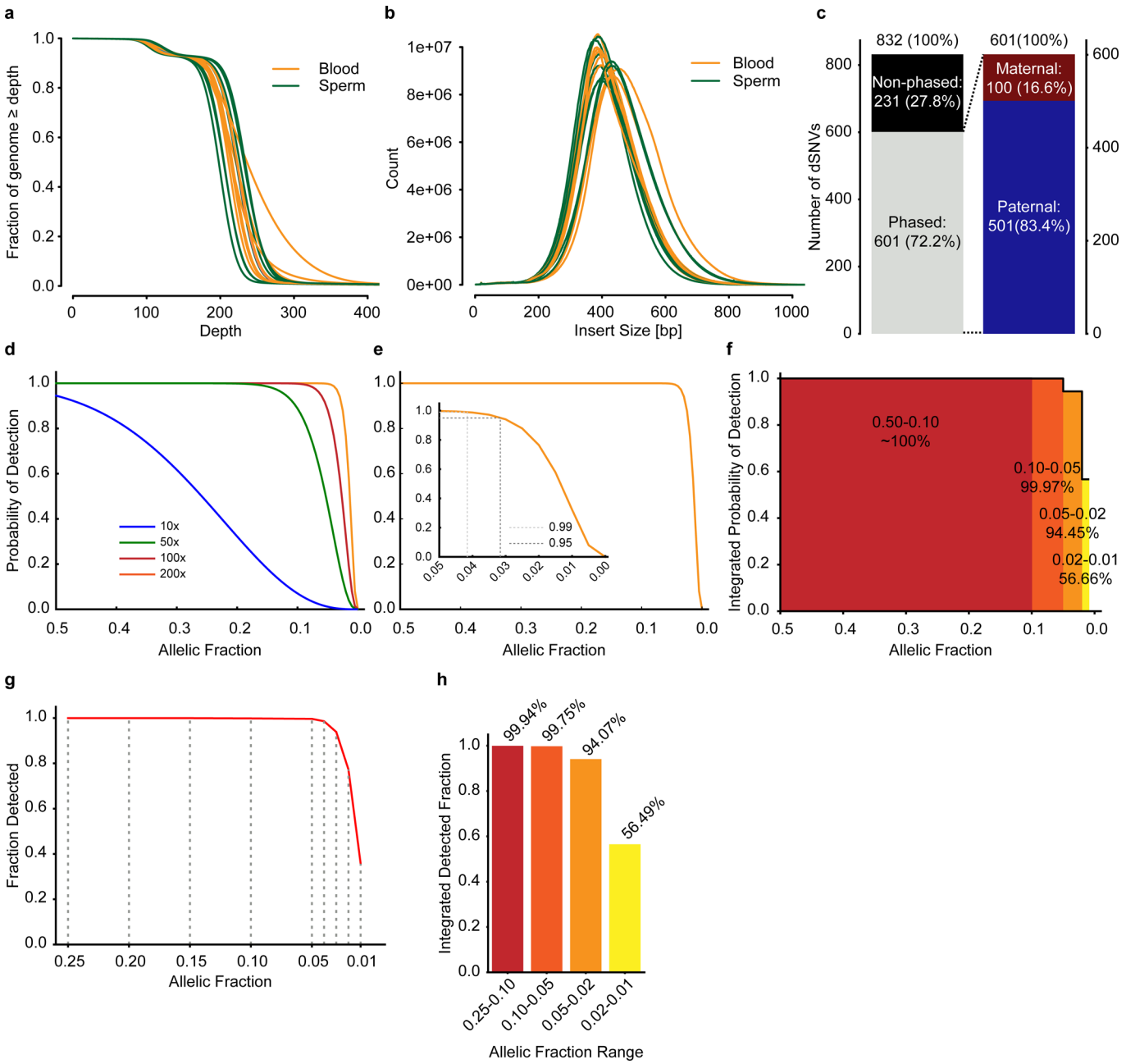
Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0711-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0711-0.

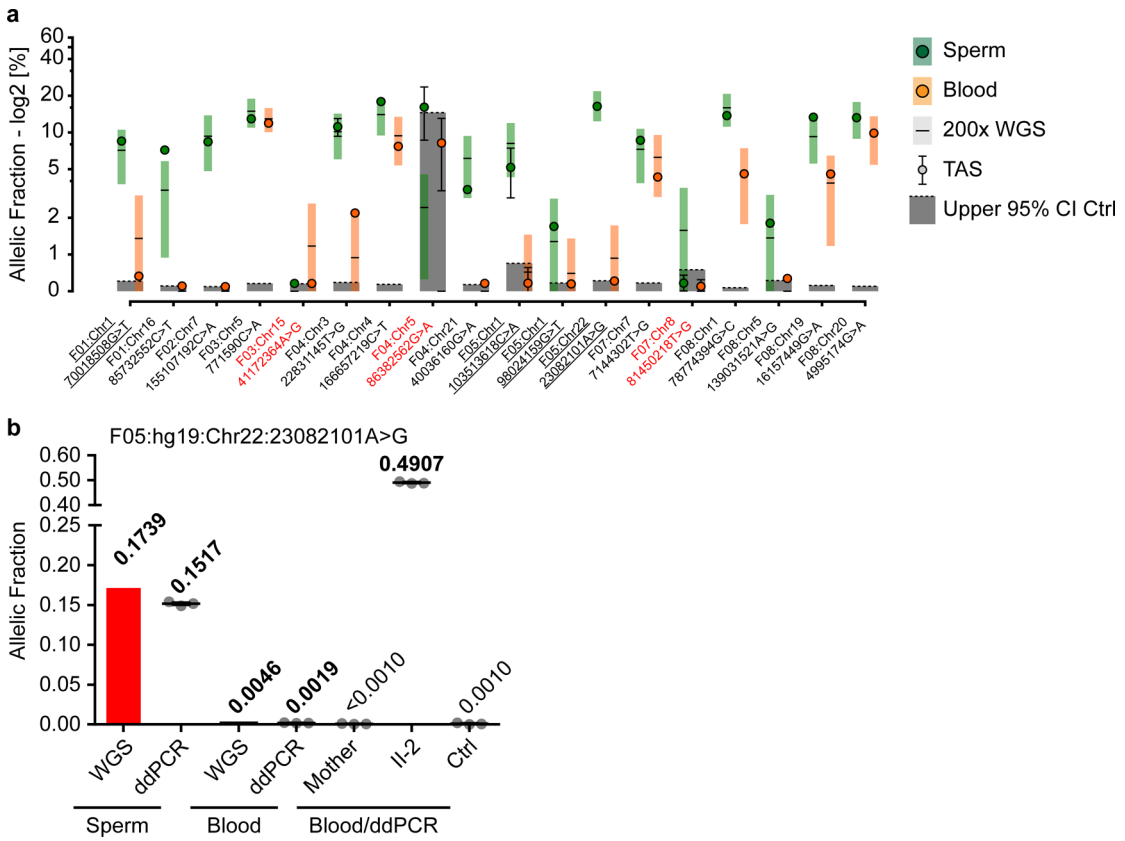
Correspondence and requests for materials should be addressed to J.S. or J.G.G.

Peer review information Kate Gao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

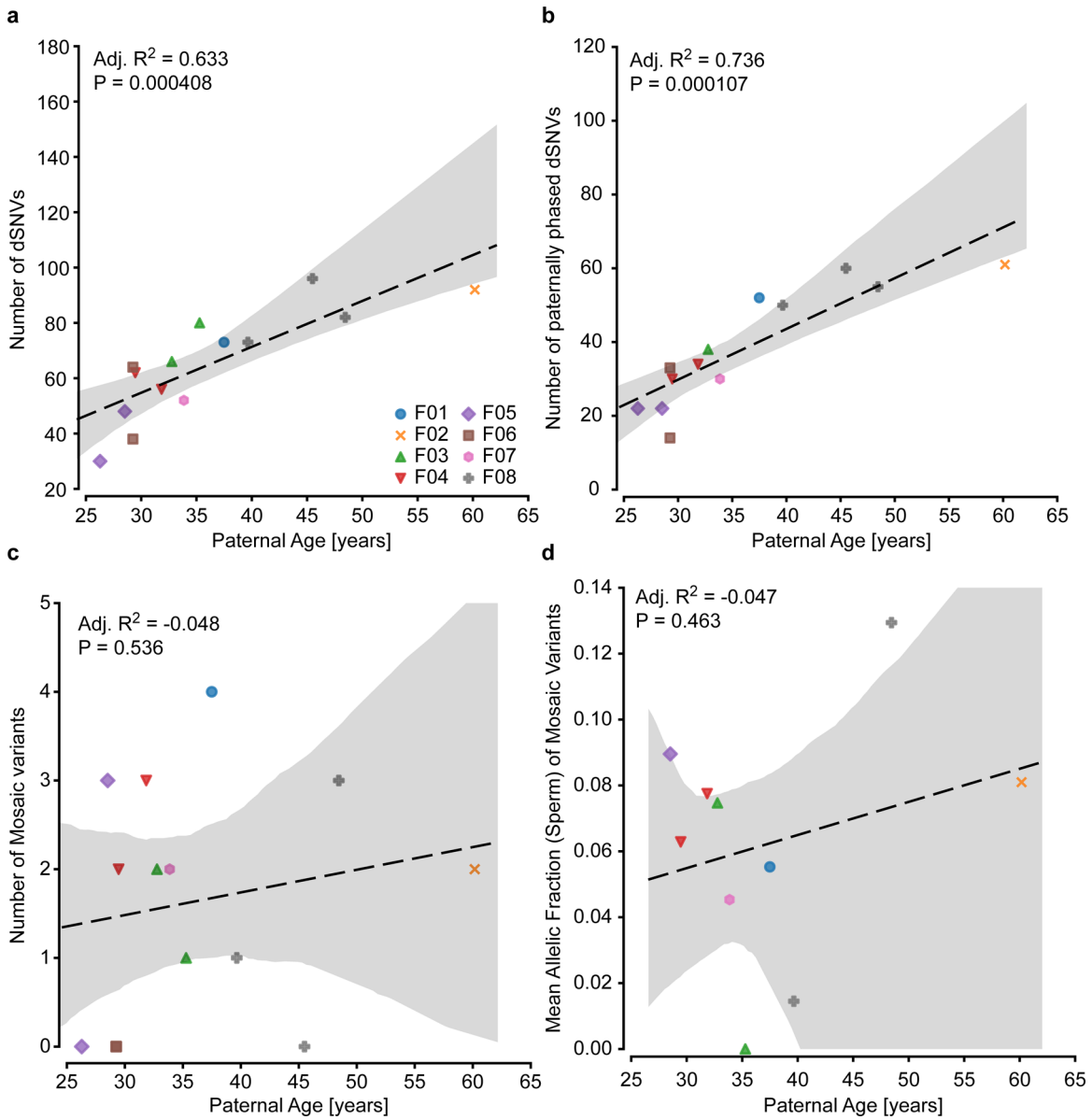
Reprints and permissions information is available at www.nature.com/reprints.



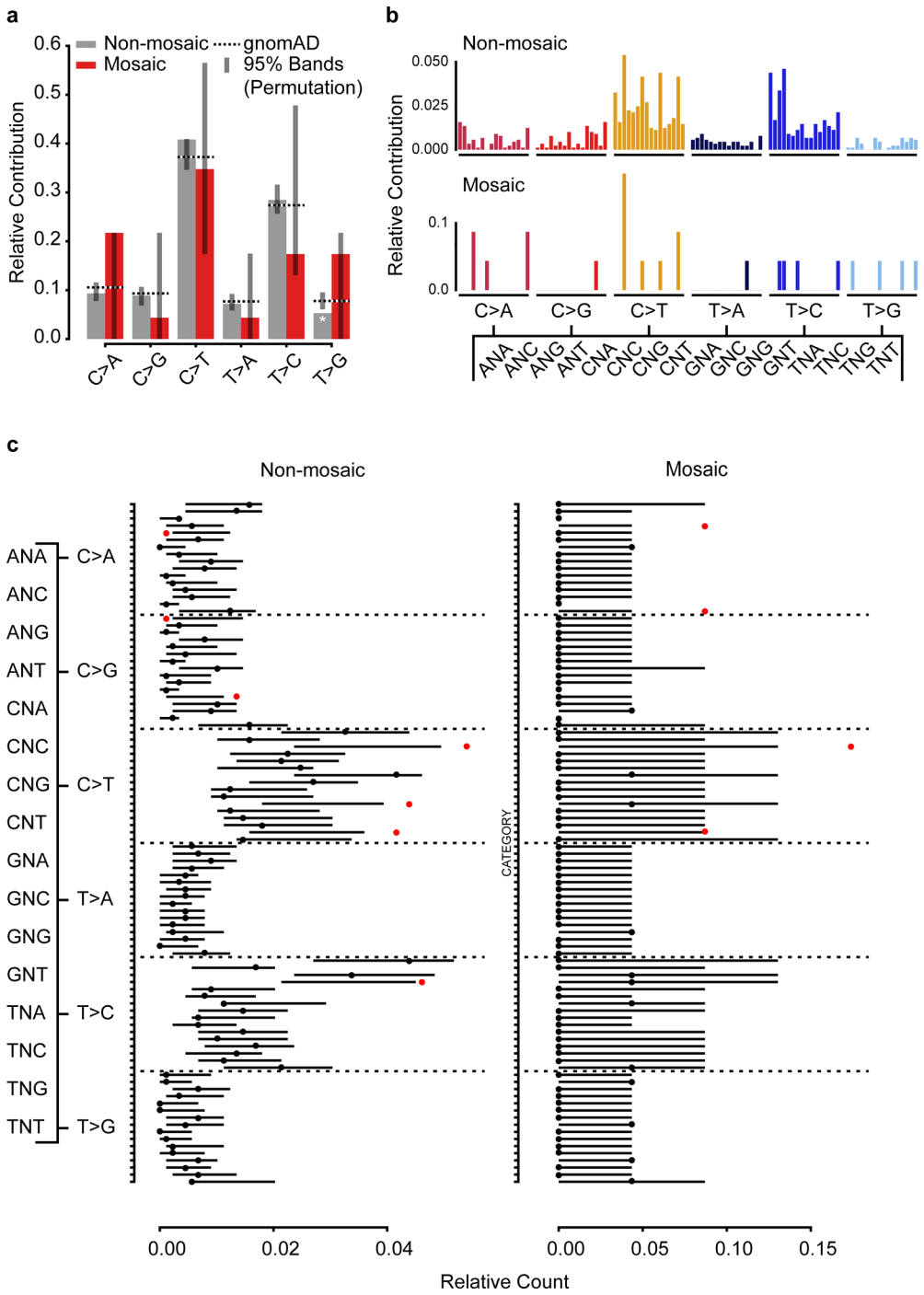
Extended Data Fig. 1 | 200×WGS allows detection of mosaic variants down to 1% sensitivity. a, Plot showing the fraction of the genome that is covered at a given depth for blood and sperm following WGS with a target coverage of 200×. **b**, Plot showing the insert size of the reads for blood and sperm. **c**, Nanopore long-read technology (average read length 5,349 bp) was able to assign parental haplotype to 601/832 dSNVs in 13 children. Out of these, 501 were paternal, resulting in α~4 as reported previously. **d-e**, Binomial models for the detection limit of mosaic variants. Plots show the probability of detecting a given variant at a specific allelic fraction (AF) when requiring at least 3 alternate reads at different read-depths (**d**) or including a magnified inset for AF between 0.05 and 0 at 200× (**e**). **f**, Analysis of the power of detection assuming a minimum requirement of 3 reads at 200× sequencing. Plot shows the integrated probability of detection for the indicated tiers based on the curve seen in **e**. **g-h**, Plot of the fraction of detected variants (**g**) and the integrated detected fraction for the indicated AF ranges (**h**) of simulated data using Pysim. Results are from 10,000 variants simulated at 0.25, 0.20, 0.15, 0.10, 0.05, 0.02, and 0.01 AF. HaplotypeCaller was employed to detect variants as for data in Fig. 1.



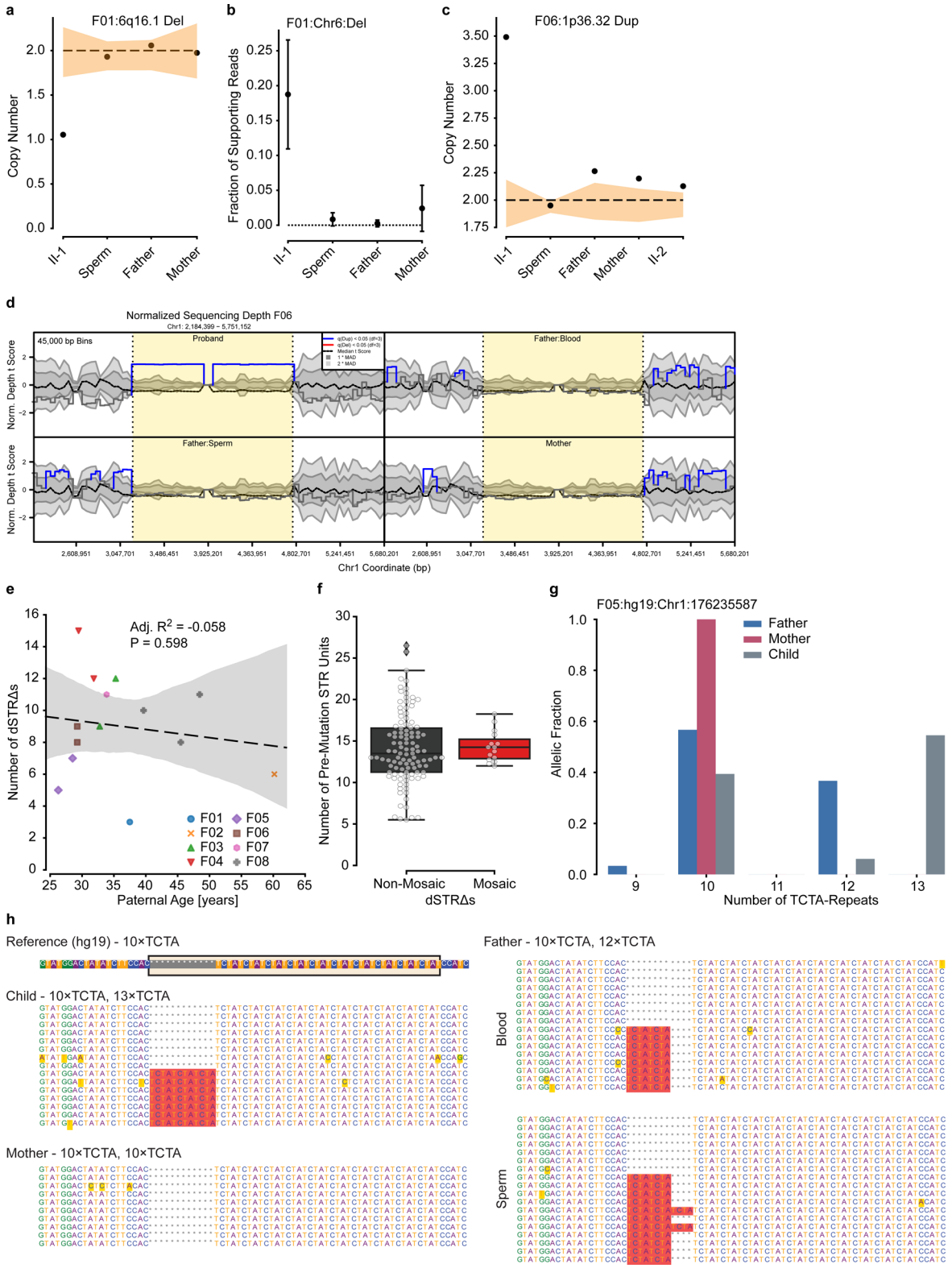
Extended Data Fig. 2 | Orthogonal validation of a subset of mosaic dSNVs. a, 18 variants that could be assessed by ultra-deep target amplicon sequencing (TAS): shown are the reported 200 × WGS results (square with horizontal line) and the results from TAS (closed circle) (shown are estimated fraction ± binomial 95% CI). Sperm (left, green) and blood (right, orange). Dashed line and grey box: upper 95% CI of an unrelated control and the area beneath to visualize likely false positive variants. y-axis: allelic fraction (%) for a log2 transformation of the data. Red text: variants that were considered to have failed orthogonal validation: 15/18 variants were successfully confirmed. Underlined variants were confirmed, but likely annotated as the wrong class (all 5 are probably SDO rather than SDE). For all data points, the estimated fraction and CI are based on the fraction of mutant reads, see Supplementary Data 2 and 4. **b**, Allelic fraction (determined by ddPCR or WGS read counts) of the mutant allele with the highest allelic fraction in sperm (F05: Chr22:23082101A > G). Sperm and Blood indicate samples from the father, other samples (Blood/ddPCR) were derived from the mother, the child harboring the dSNV (II-2), or control (Ctrl) blood. Graph shows individual data points (experimental triplicates) and mean ± SEM for the ddPCR data.



Extended Data Fig. 3 | Age correlation of all and mosaic dSNVs. a, Plot showing the increase in dSNV number with paternal age at birth, as described previously[1,5]. Dashed line shows a regression curve demonstrating this dependence (n = 14 trios, adjusted R^2 = 0.526, P = 0.0020). **b**, Plot showing the increase in dSNV number with paternal age at birth for paternal variants only. As expected, this correlation was stronger than for non-phased variants (n = 13 trios, adjusted R^2 = 0.736, P = 0.000107). **c-d**, Plots showing correlation for paternal age and the number of mosaic variants or the mean AF in sperm. *Paternal age/the number of mosaic variants* (**c**; n = 14 trios, adjusted R^2 = -0.048, P = 0.536) and *paternal age/mean AF in sperm* (**d**; n = 14 trios, adjusted R^2 = -0.047, P = 0.463) did not show any significant correlation. Adjusted R^2, coefficient of determination, and F-statistic nominal P-values are derived from a linear regression model through ordinary least squares. All graphs show individual data points, a regression line, and the 95% CI.

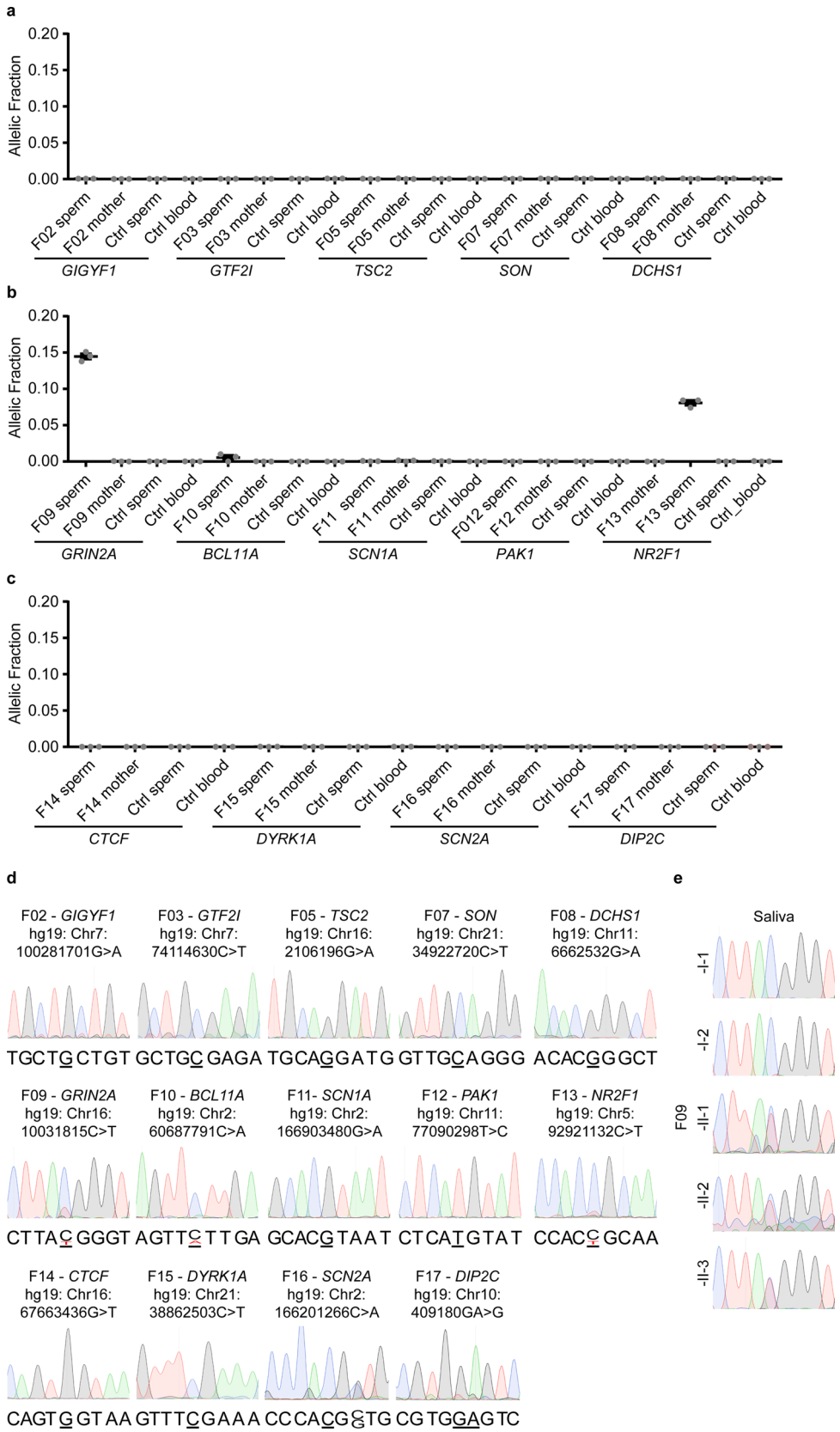


Extended Data Fig. 4 | Mutational signature for non-mosaic and mosaic dSNVs. a, Mutational signatures (6 categories) for non-mosaic and mosaic dSNVs, compared to the overall gnomAD signature and a permuted subset (n = 1,000 permutations for n = 889 (non-mosaic) and n = 23 (mosaic) dSNVs; shown is the 95% band). Asterisks indicate observed signatures that lie outside the 95% band of the permuted variants.. Non-mosaic variants are largely reminiscent of the gnomAD signature (with the exception of a significant depletion of T > G). Mosaic variants exhibit some differences, but none reach significance due to the low number of available mutations. **b**, Mutational signatures (96 categories; trinucleotide environment for non-mosaic and mosaic dSNVs. **c**, Detailed view of the 96 mutational categories for non-mosaic and mosaic dSNVs, compared to the overall gnomAD signature and a permuted subset (n = 1,000 permutations for n = 889 (non-mosaic) and n = 23 (mosaic) dSNVs; shown is the 95% band). Dots indicate the observed mutational signature (black: within 95% band; red: outside the 95% band).

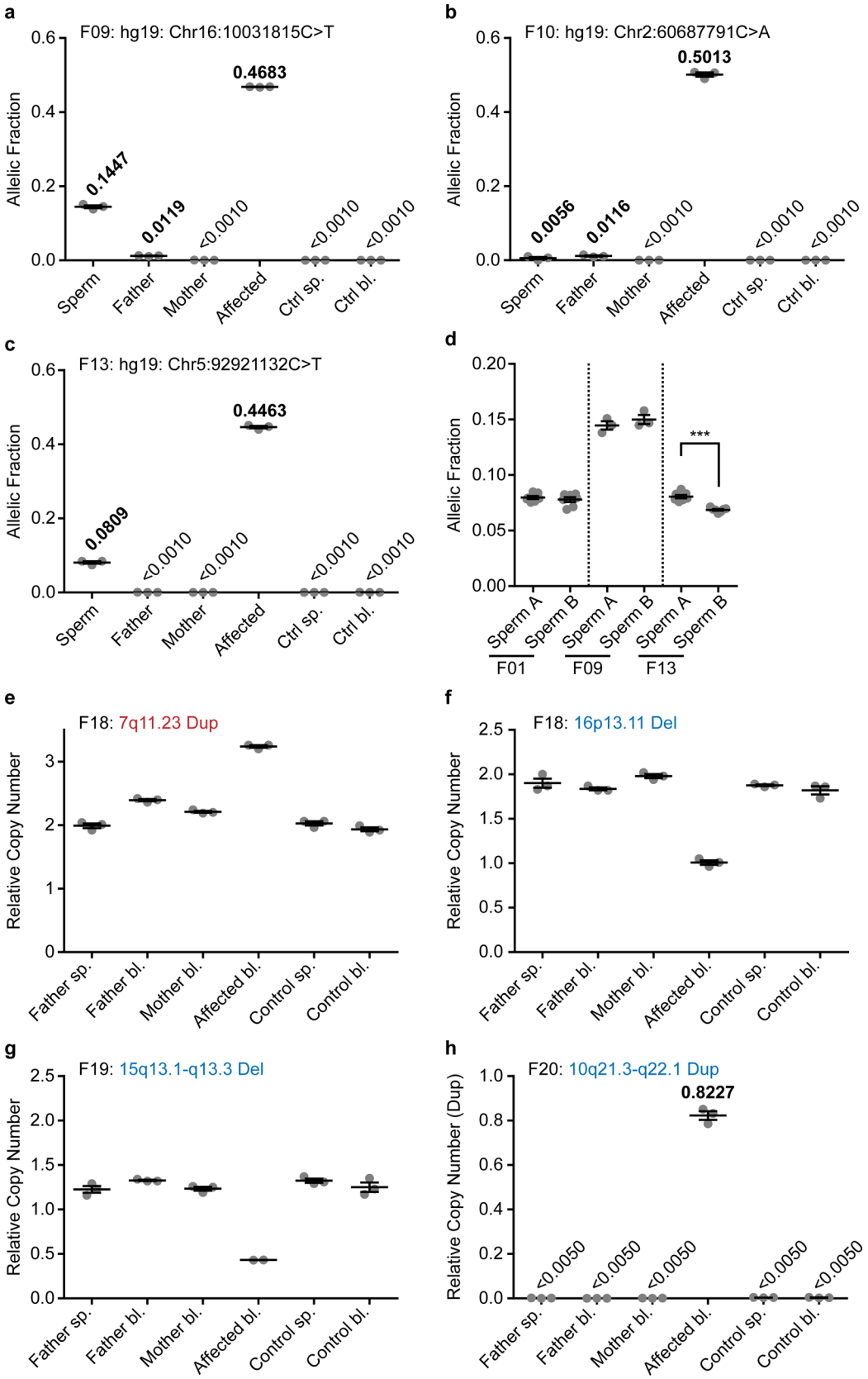


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Sperm mosaicism stratifies recurrence risk for dSV and dSTRΔ variants. a-c, Calculated copy number (**a**, **c**) and fraction of supporting reads (**b**) for the 6q16.1 deletion in F01 and The 1p36.32 duplication as indicated. Orange band in **a** and **c**: ±1 SD of the CN using similarly sized regions across the genome (n = 1,000 random regions, see Methods). Plot in **b** shows the estimated fraction of supporting reads (estimated fraction ± binomial 95% CI; based on the fraction of mutant reads, see Supplementary Data 7). Together, these approaches suggest that these dSVs are not mosaic in paternal sperm. Note that the fraction of supporting reads could not be used for the duplication due to the repetitive elements flanking this SV. **d**, Copy number variant plot for the duplication in F06 for the Proband (40×), Father (200× both), and the mother (40×). Visualization was performed with the CNView[36] tool (see Methods). **e**, Correlation of the number of dSTRΔs with paternal age at birth. Dashed line shows a regression curve (n = 14 trios, adjusted R^2 = -0.058, P = 0.598). Adjusted R^2, coefficient of determination, and F-statistic nominal P-value are derived from a linear regression model through ordinary least squares. Graph shows individual data points, a regression line, and the 95% CI. **f**, Number of STR repeat units for non-mosaic dSTRΔs or those that are mosaic. No significant difference can be observed between the two groups (n = 111 non-mosaic variants and n = 15 mosaic variants; two-tailed Mann Whitney test; nominal P = 0.5490). Boxplots show median and quartiles with outliers as well as individual values. **g**, Detailed analysis of the TCTA repeat numbers in paternal, maternal, and child's blood at low sequencing depth. Results show a de novo 13× repeat in the child that is neither present in the father nor the mother. **h**, Sample reads showing the presence of a 10× and 13× allele in the child, a homozygous 10× allele in the mother, a 10× and a 12× allele in the father, and the presence of a mosaic 13× allele exclusively in paternal sperm.

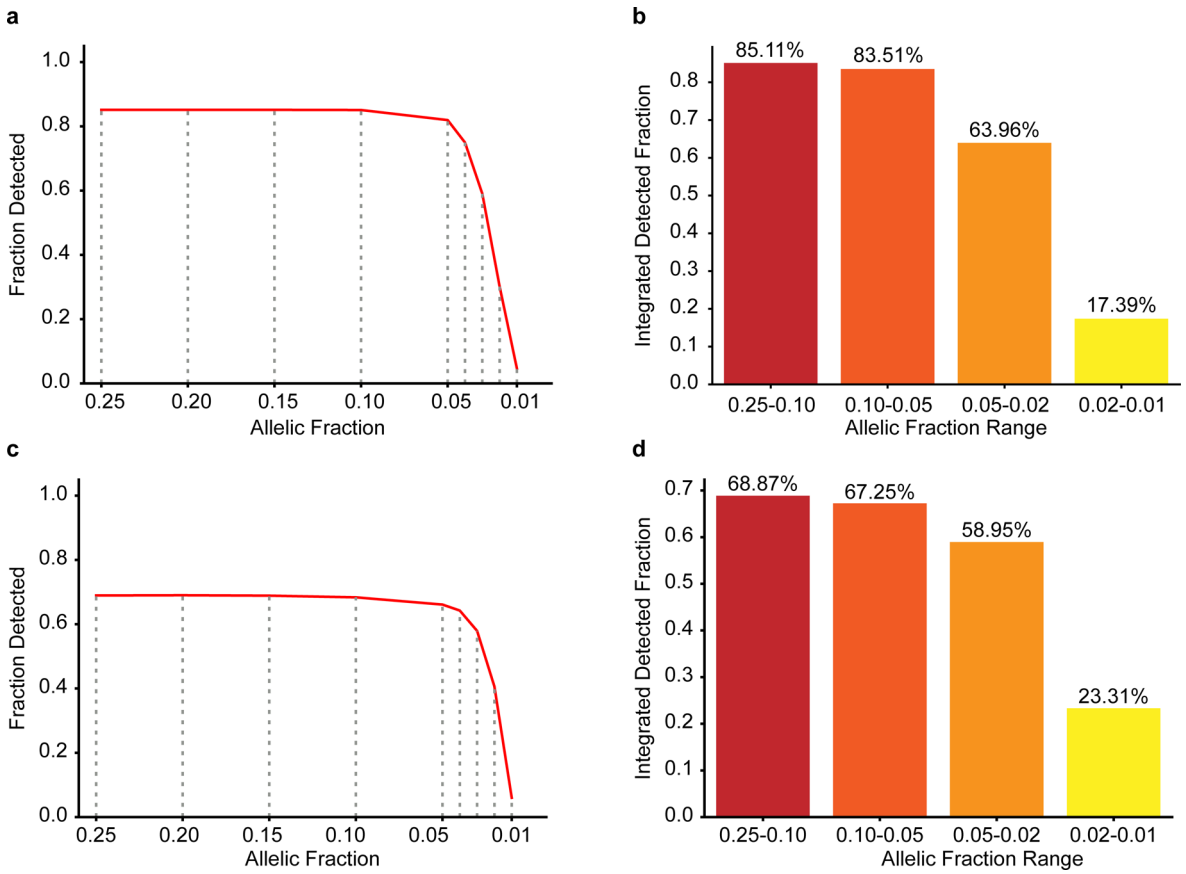


Extended Data Fig. 6 | Sperm mosaicism stratifies risk for pathogenic ASD mutations. a-c AF (determined by ddPCR) of the mutant allele in paternal sperm (sperm) and maternal blood (mother) for the relevant dSNV in the 14 families. Part of this panel is also presented in Fig. 3. Ctrl –an unrelated sperm or blood sample, as indicated, acting as control. Graphs show individual data points (experimental triplicates) and mean ± SEM. **d**, Sanger sequencing results of paternal sperm for the locus harboring the dSNV for each family. Confirming the ddPCR results, F09, F10, and F13 showed mosaicism at their respective positions. **e**, Sanger sequencing results showing the C > T conversion locus in *GRIN2A* in F09 for all family members. The mutation was absent in the saliva of both parents, but present as a heterozygous allele in all 3 children.

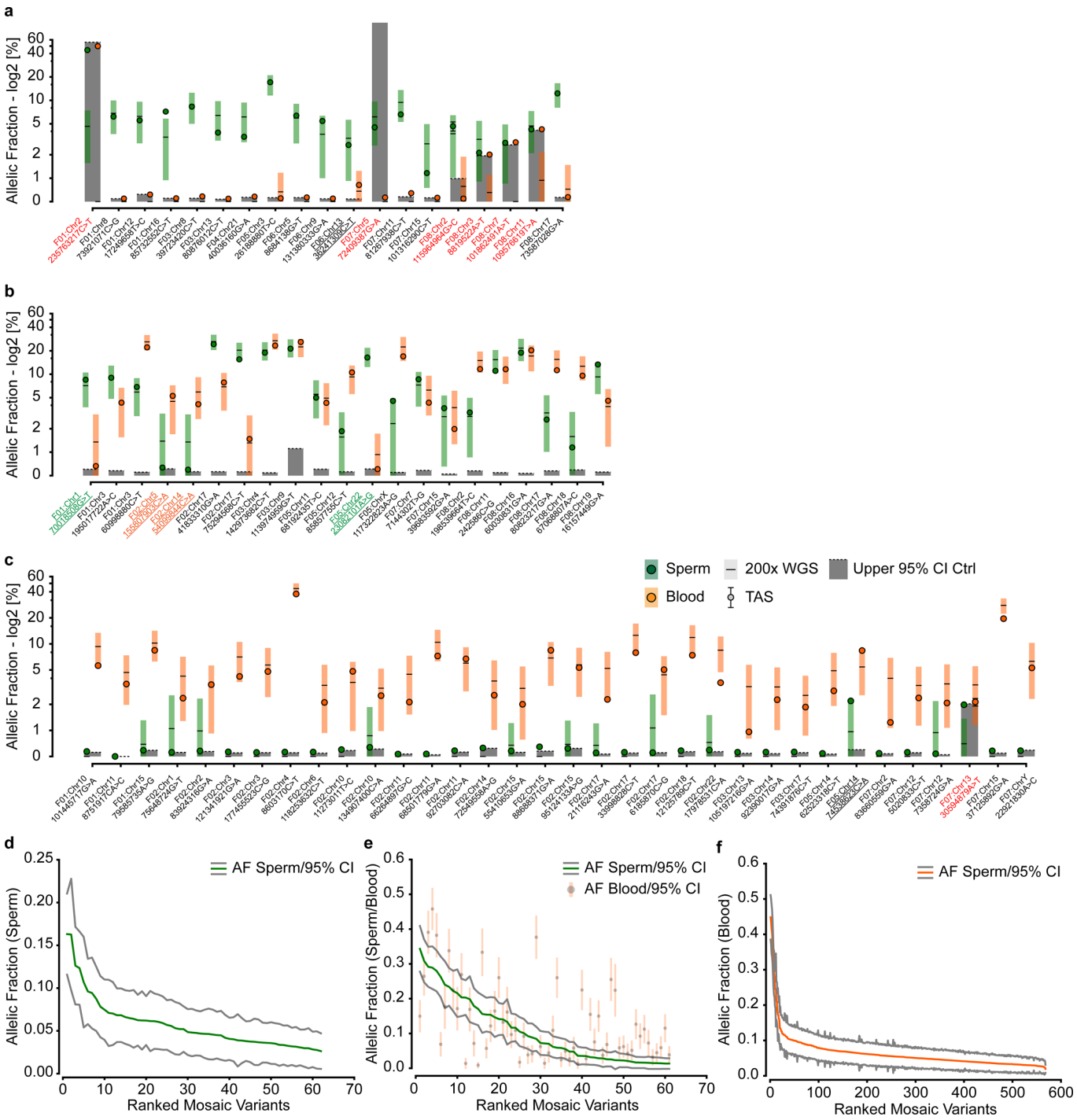


Extended Data Fig. 7 | See next page for caption.

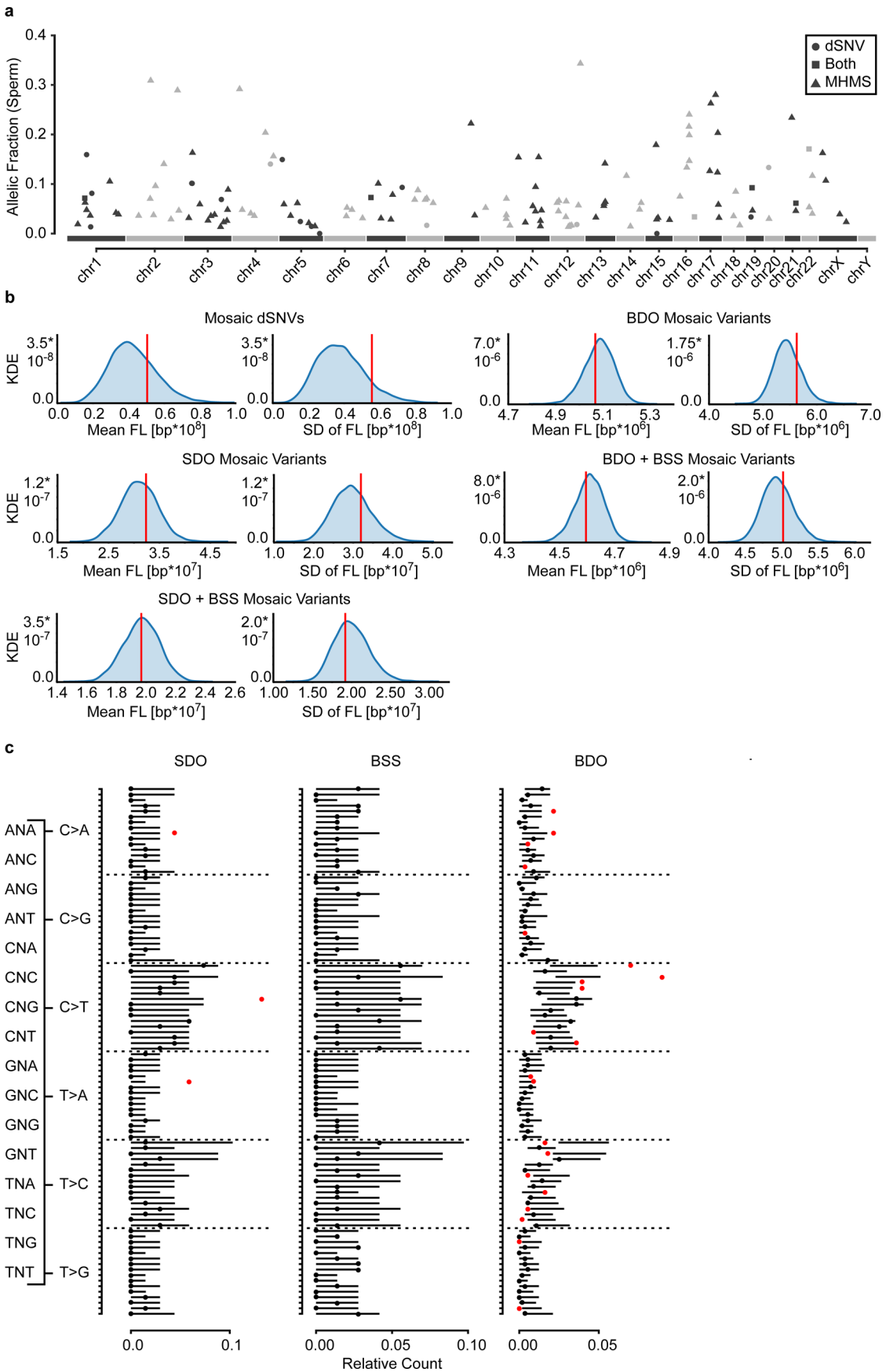
Extended Data Fig. 7 | ddPCR assessment of pathogenic structural variants and recurrent sampling of pathogenic DNMs in F01, F09, and F13. a-c, AF (determined by ddPCR) of the mutant alleles in F09 (**a**), F10 (**b**), and F13 (**c**). DNA tested was derived from paternal sperm and the saliva (**a** and **b**) or blood (**c**, bl.) of the father, mother, or affected child. In addition, controls for sperm (sp) and blood (bl) are provided. **d**, AF (determined by ddPCR) comparing two biological replicates of paternal sperm for F01, F09, and F13. The samples showed comparable levels of AF over time for all three samples, however, F13 exhibited a minor, but statistically significant difference. ***$P < 0.001$ (unpaired t-test, two-tailed, degrees of freedom = 12). **e-g**, Relative copy number (determined by ddPCR) for the three indicated dSVs for blood- and sperm-derived samples. Note that there is no detectable abnormality in the paternal sperm copy number above noise level, suggesting absence of sperm mosaicism in these samples. **h**, Direct copy number quantification of the duplication by ddPCR. All graphs show individual data points (experimental triplicates except for Affected in **g** [experimental duplicate], and F01 and F13 in **d** [7 experimental replicates]) and mean ± SEM.



Extended Data Fig. 8 | Limit of detection analysis for the unbiased analysis of gonadal mosaic SNVs. a-d, Plots of the fraction of detected variants (**a**, **c**) and the integrated detected fraction for the indicated AF ranges (**b**, **d**) of simulated data using Pysim for the intersection of MuTect 2/Strelka 2 (**a**, **b**) and MosaicHunter (**c**, **d**). Results were from 10,000 variants simulated at 0.25, 0.20, 0.15, 0.10, 0.05, 0.02, and 0.01 AF. This was the same data set as used in Extended Data Fig. 1. The MuTect 2/Strelka 2 and MosaicHunter pipelines were employed with the same filters as for the data in Fig. 4.



Extended Data Fig. 9 | Mosaic SNVs identified by unbiased analysis have a high validation rate and their AF differs depending on their origin. a-c, 74 variants that could be assessed by ultra-deep target amplicon sequencing (TAS): shown are the reported 200 × WGS results (square with horizontal line) and the results from TAS (closed circle) (shown are estimated fraction ± binomial 95% CI). Sperm (left, green) and blood (right, orange). Dashed line and grey box: upper 95% CI of an unrelated control and the area beneath to visualize likely false positive variants. y-axis: allelic fraction (%) for a log2 transformation of the data. Plots are split by the three categories: SDO (**a**), BSS (**b**), and BDO (**c**). Red text denotes variants that were considered to have failed orthogonal validation: 13/19 (**a**), 21/21 (**b**), and 33/34 (**c**) were successfully confirmed. Underlined variants were confirmed, but likely annotated as the wrong class (that is, they are actually BSS for SDO and BDO variants in **a** and **c**, or are SDO (green text) or BDO (orange text) for BSS variants in **c**). For all data points, the estimated fraction and CI are based on the fraction of mutant reads, see Supplementary Data 2 and 8. **d-f**, Ranked plot of the estimated sperm and blood AF with 95% confidence intervals (estimated fraction ± binomial CI; based on the fraction of mutant reads, see Supplementary Data 8) for all variants detected in the three categories. SDO (**d**) and BDO (**f**) variants both show curves that are reminiscent of exponential decay, consistent with an increase of the number of mutations with expansion of the progenitor pool at a constant mutational rate. However, BSS (**e**) mosaicism for the first 40 variants appears to be more linear, suggesting that mutation rates for early division might be higher than those for later. This is consistent with previous models that estimated an elevated mutation rate in early embryonic development[14].



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Mosaic variants do not exhibit clustering but differ in their mutational signatures depending on their origin. a, Plot of the chromosomal location for each of the mosaic variants and their allelic fraction found in sperm from F01-08. Circles, triangles, and squares denote variants found to be mosaic by the dSNV approach, by the unbiased approach, or by both, respectively. **b**, Permutation simulations (n = 10,000 simulations of n = 23 mosaic dSNVs, n = 62 SDO mosaics, n = 123 SDO + BSS mosaics, n = 568 BDO mosaics, and n = 629 BDO + BSS mosaics) of variant locations to obtain mean and SD of broken stick fragment lengths. Vertical lines mark the observed value from mosaic dSNVs and mosaic variants from the indicated classes. These simulations illustrate that the observed distributions of variants along the chromosomes (as visualized in **a** for those that were mosaic in sperm) were within expectation. **c**, Detailed view of the 96 mutational categories for SDO, shared, and BDO mosaic variants,, compared to the overall gnomAD signature and a permuted subset (n = 1,000 permutations for n = 68 (SDO), 72 (BSS), and 568 (BDO) gnomAD SNVs; shown is the 95% band). Dots indicate the observed mutational signature (black: within 95% band; red: outside the 95% band).

nature research

Corresponding author(s): Joseph G. Gleeson, Jonathan Sebat

Last updated by author(s): Nov 15, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data in this study.
Data analysis	Commercial and available software used directly to analyze and visualize data: QuantaSoft (1.7.4.0917) and QuantaSoft Analysis Pro (1.0.596), GraphPad Prism 5, R (3.4.1 and ggbio package), Python (3.6.5 and matplotlib, seaborn, pysam, pandas, scipy modules), CNView, Strelka 2 (v2.9.2), MuTect 2 (v2.1), MosaicHunter (v1.0), HipSTR (v0.6), Pysim, Picard's MarkDuplicates (v1.83), BWA (v0.7.8), Genome Analysis ToolKit (GATK version 3.5-0-g36282e4), SnpEff (v4.2), SnpSift (v4.2), Triodenovo (v0.06), BWA mem (version 0.7.15-r1140), sambamba (version 0.6.6), samtools (v1.9), and bedtools (v2.25.0). Variant calling was done using standard pipelines and programs as described in the methods section of this paper.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during the current study will be made available in public data bases and are available upon request from J.G.G. on reasonable request. Additionally, summary tables of all data are included as supplementary information files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for deep sequencing was based on previous mosaic studies on blood and the expected number of de novo mutations per trio. Based on this, we estimated that we require at least 10 trios to obtain a minimum of 500 DNMs for interrogation (current n=14 trios with n=912 dSNVs), to obtain >10 mosaic variants for analysis. Sequencing depth was determined by theoretical considerations (binomial model) and simulated data. For the analysis of pathogenic DNMs, based on previous sperm mosaicism studies on epileptic disorders, we estimated that around 10% of pathogenic variants may be mosaic. Thus, we expected that we need at least 15 DNM families to observe a mosaic variant with ~80% chance (n=20 pathogenic DNMs).
Data exclusions	No full data sets were excluded in this study. Individual variants were filtered based on previously established best practices for variant calling and mosaicism detection.
Replication	This is a descriptive study of fixed, unique cohort. Replication was not attempted.
Randomization	This is a descriptive study. No randomization was performed.
Blinding	No groups were allocated by the scientists.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human participants were derived from two autism cohorts (O.D. and J.S.). The study included parents, affected children and unaffected children of all sexes and ages.
Recruitment	The sole inclusion criterion was a medical diagnosis of autism spectrum disorder with or without epilepsy and a molecular diagnosis by next generation sequencing. The patient cohort was solely used for genetic analysis and the ascertainment of appropriate samples was performed as stated in our IRB protocol. In short, fathers of individuals with autism were previously enrolled through independent studies and subjected to whole genome or exome sequencing. All fathers with a pathogenic DNM who agreed to be recontacted for additional studies were informed of this study. As the existence of sperm mosaicism is independent of any possible selection bias (e.g. religious conflicts, history of vasectomy), willingness to participate and provide a semen sample is expected to be independent from measured outcomes.
Ethics oversight	IRB at UC, San Diego

Note that full information on the approval of the study protocol must also be provided in the manuscript.