Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder

Elaine T Lim1–4 , Mohammed Uddin5, Silvia De Rubeis6,7, Yingleong Chan3,4, Anne S Kamumbu1–3, Xiaochang Zhang1–3, Alissa M D'Gama1–3, Sonia N Kim1–3, Robert Sean Hill1–3, Arthur P Goldberg6,7, Christopher Poultney6,7, Nancy J Minshew8, Itaru Kushima9, Branko Aleksic9 , Norio Ozaki9, Mara Parellada10, Celso Arango10, Maria J Penzol11, Angel Carracedo12–14, Alexander Kolevzon6,7,15–17, Christina M Hultman18, Lauren A Weiss19, Menachem Fromer6,7,20, Andreas G Chiocchetti21, Christine M Freitag21, Autism Sequencing Consortium22, George M Church3,4, Stephen W Scherer23–26, Joseph D Buxbaum6,7,15,27 & Christopher A Walsh1–3

1 Division of Genetics and Genomics, Manton Center for Orphan Disease Research and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, Massachusetts, USA.

2 Departments of Pediatrics and Neurology, Harvard Medical School, Boston, Massachusetts, USA.

3 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

4 Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA.

5 Mohammed Bin Rashid University of Medicine and Health Sciences, College of Medicine, Dubai, United Arab Emirates.

6 Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

7 Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

8 Department of Psychiatry, Center For Excellence in Autism Research, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

9 Department of Psychiatry, Nagoya University Graduate School of Medicine, Nagoya, Japan.

10 Child and Adolescent Psychiatry Department, Hospital General Universitario Gregorio Marañón, School of Medicine, Universidad Complutense, IiSGM, CIBERSAM, Madrid, Spain.

11 Child and Adolescent Psychiatry Department, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, Madrid, Spain.

12 Grupo de Medicina Xenomica, Universidade de Santiago de Compostela, Centro Nacional de Genotipado-Plataforma de Recursos Biomoleculares y Bioinformaticos-Instituto de Salud Carlos III (CeGen-PRB2-ISCIII), Santiago de Compostela, Spain.

13 Grupo de Medicina Xenomica, CIBERER, Fundacion Publica Galega de Medicina Xenomica-SERGAS, Santiago de Compostela, Spain.

14Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

15 Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

16 The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

17 Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

18 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

19 Department of Psychiatry and Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA.

20 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

21 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Autism Research and Intervention Center of Excellence, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany.

22 A list of members and affiliations appears in the Supplementary Note.

23 The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada. 24 Program in Genetics and Genome Biology (GGB), The Hospital for Sick Children, Toronto, Ontario, Canada.

25 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

26 McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada.

27 The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Correspondence should be addressed to C.A.W. (Christopher.Walsh@childrens.harvard.edu) or E.T.L (elimtt@gmail.com).

We systematically analyzed postzygotic mutations (PZMs) in whole-exome sequences from the largest collection of trios (5,947) with autism spectrum disorder (ASD) available, including 282 unpublished trios, and performed resequencing using multiple independent technologies. We identified 7.5% of de novo mutations as PZMs, 83.3% of which were not described in previous studies. Damaging, nonsynonymous PZMs within critical exons of prenatally expressed genes were more common in ASD probands than controls ($P < 1 \times 10-6$), and genes carrying these PZMs were enriched for expression in the amygdala ($P = 5.4 \times 10-3$). Two genes (KLF16 and MSANTD2) were significantly enriched for PZMs genome-wide, and other PZMs involved genes (SCN2A, HNRNPU and SMARCA4) whose mutation is known to cause ASD or other neurodevelopmental disorders. PZMs constitute a significant proportion of de novo mutations and contribute importantly to ASD risk.

ASD is a complex disorder with genetic and clinical heterogeneity. Beyond common variation1, previous studies focusing on germline mutations have demonstrated a significant contribution from de novo copy number variants (CNVs)2,3, and more recent whole-exome sequencing (WES) analyses have highlighted the role of de novo point mutations4,5. Although the number of exonic de novo mutations is similar between affected and unaffected individuals (~1 de novo point mutation per exome), ASD probands harbor an excess of deleterious and loss-of-function (LoF) de novo mutations in exons compared to their unaffected siblings4,5. Collectively, 4–7% of probands have a de novo CNV and ~7% of probands have a de novo point mutation that confers risk to ASD2. Additionally, WES studies have uncovered increased ASD risk from rare autosomal recessive (3%) and X-linked variants (2%)6,7. However, a large portion of ASD risk cannot be explained by germline de novo,

recessive and X-linked variants, and this warrants investigation of other genetic contributions to ASD risk.

Since PZMs arise after fertilization, they result in distinct cell populations within the same individual that can contribute to varying disease manifestations. These mutations are typically not transmitted to offspring, and it has been hypothesized that PZMs account for a significant proportion of genetic risk in sporadic disorders. There is increasing recent evidence that PZMs can contribute to brain malformations and epilepsy8,9 and that a fraction of clinically relevant PZMs can be detected in blood of affected individuals8,10. The role of PZMs in ASD risk is unknown, and we therefore explored the contribution of this type of variation to ASD. PZMs are efficiently detected by candidate-gene sequencing panels, given their deep sequencing cover-age. However, PZMs present in greater than 25–30% of cells (or 15% alternate allele fraction (AAF)) can be detected with reasonable sensitivity using WES8. We recalled WES data from 5,947 trios, adding 282 newly sequenced trios from the Autism Sequencing Consortium and Simons Simplex Collection, and, using a custom pipeline, we resequenced PZMs detected from WES data using three resequencing technologies, providing a systematic evaluation of PZMs' contribution to ASD risk.

RESULTS

Excess de novo mutations with low AAFs

We analyzed de novo mutations in WES data from 5,947 families, which included 4,032 ASD trios and 1,918 quads that also have unaffected siblings (Supplementary Tables 1–3)4,5. The vast majority of samples (96%) were derived from whole-blood DNA, and a negligible fraction was derived from lymphoblastoid cells (3%) and primary saliva (1%). We included all samples derived from various tissue types but removed outlier samples with a large number of de novo or mosaic mutations from our analyses (Online Methods). We increased specificity for likely pathogenic mutations by filtering out variants that were present in control exomes, resulting in modestly lower rates of de novo mutations than previously called (4,846 in total). Of these, a substantial portion (23%) showed low AAFs of ≤ 40% (Fig. 1a). The modal AAF was ~50%, which is consistent with the expected AAF for a germline heterozygous mutation. We observed a 1.4-fold excess of mutations in the 40–50% AAF category, compared to the 50–60% AAF category, suggesting a modest bias toward mutations with lower AAFs, possibly due to amplification, capture or sequencing biases for the alternate alleles. In contrast, we observed a robust (4.1-fold) excess of mutations with AAF ≤ 40% (23.7% of all de novo mutations), compared to those with AAF ≥ 60% (5.8% of all de novo mutations), suggesting that a substantial proportion of mutations with AAF ≤ 40% arose from a biological mechanism rather than a technical bias. In addition, we found an excess of de novo point mutations compared to inherited variants in the AAF 40% category (odds ratio (OR) = 1.67), which was not seen in the AAF ≥ 60% category (OR = 0.82). This suggests that a substantial portion of de novo mutations is likely to have arisen postzygotically rather than in the parental gametes.

Detection of PZMs from WES and secondary resequencing

Given our initial observations that some de novo mutations might be PZMs, we developed a pipeline to quantitatively categorize PZMs with high or low confidence in our cohort of 5,947 ASD families (Online Methods). Of 4,846 total de novo mutations (which we define as Group A; Fig. 1b), 1,113 were candidate PZMs (23%, Group B), defined as having an AAF that was ≤ 80% of the modal AAFs, which ranged from 40–50%. Of these Group B mutations, 468 were interpreted as high-confidence PZMs (9.7%, Group C) because they showed statistically significant deviation from the modal AAFs.

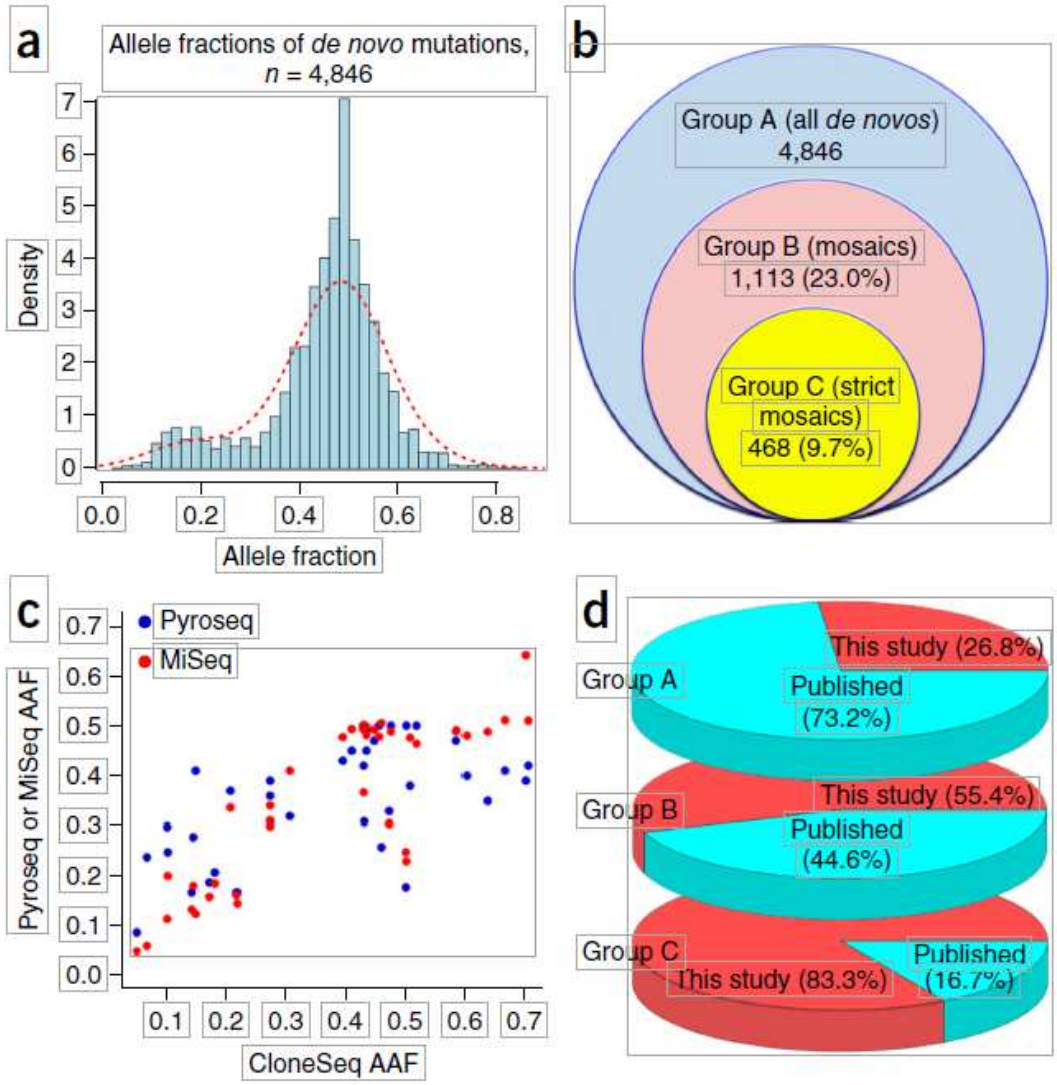

Figure 1 De novo mutations in ASD show an excess of low AAFs, consistent with postzygotic mosaicism. (a) There is an excess of variants with low AAFs among the de novo mutations, which are likely to be postzygotic mutations. (b) Rates of mutations in the data sets for all de novo mutations in Group A, as well as mosaics in Groups B and C.

(c)AAFs of PZMs detected using multiple resequencing technologies (n = 49 mutations for CloneSeq, n = 46 mutations for Pyroseq, n = 42 mutations for MiSeq), with higher correlations observed (Pearson's r2 = 0.85 for CloneSeq and MiSeq, r2 = 0.63 for CloneSeq and Pyroseq).

(d)Percentages of identified de novo variants that were identified by previous analyses or were newly identified from Groups A, B and C. The majority of high-confidence PZMs from Group C were not detected by previous calling algorithms.

We compared the 4,846 de novo mutations in Group A from our study with previous studies reporting these datasets4,5 and found that 1,297 of the de novo mutations (26.8%) we identified had not been previously reported in ASD. We enriched our set for de novo mutations that are most likely to be pathogenic and to have a large effect in ASD by filtering out de novo mutations found in control individuals. As such, our reported rate of de novo mutations is conservative: on average, less than 1 de novo mutation per exome. However, we also recalled the exomes jointly using the latest

GATK variant-calling pipelines and best practices. These practices likely account for our improved detection of previously unreported mutations.

To experimentally test the candidate PZMs, we applied independent resequencing methods in three phases. In Phase 1, we resequenced 50 mutations, based on sample availability, across the three groups (Table 1) using three independent technologies—pyrosequencing (Pyroseq), subcloning with Sanger colony sequencing (CloneSeq) and targeted PCR followed by MiSeq resequencing (Supplementary Table 4)—to test whether these mutations deviated from the expected AAF of 50% and to compare these technologies. We found that 84.8–93.3% of the Group C mutations, predicted to be high-confidence PZMs, were indeed likely to arise postzygotically with confirmed AAFs ≤ 40% (Table 1). Of the less stringent candidate PZMs (in Group B but not in C), 25–38% were confirmed as postzygotic with AAFs ≤ 40%. In Phase 2, we resequenced another 181 mutations from all groups using targeted PCR and MiSeq, as well as Pyroseq, and replicated the rates observed in Phase 1: 84.8–85.2% of high-confidence PZMs (Group C) and 13.5–25.6% of less stringent PZMs (Group B) showed AAFs ≤ 40%. A small percentage (8.3%) of predicted germline de novo mutations (gDNMs) found only in Group A (and not identified as also being in Groups B or C) also showed AAFs ≤ 40%. In Phase 3, we resequenced 325 mutations using targeted PCR and MiSeq with DNA derived from blood samples and found that 97% of high-confidence PZMs, 17.6% of less stringent PZMs and 2.8% of predicted gDNMs have AAFs ≤ 40%.

Table 1 Validation rates for mutations detected from WES

	High-confidence PZMs from Group C	Less-stringent PZMs found in Group B but not Group C	Potential gDNMs found in Group A but not Group B
Phase 1: Resequencing of initial 50 mutations to evaluate whether AAFs≤ 40%			
CloneSeq	14 of 16 (87.5%)	7 of 28 (25%)	1 of 5 (20%)
Pyroseq	13 of 15 (87%)	10 of 26 (38%)	2 of 5 (40%)
Targeted PCR + MiSeq	14 of 15 (93.3%)	6 of 24 (25%)	0 of 3 (0%)
Phase 2: Resequencing of 181 mutations to evaluate whether AAFs≤ 40%			
Pyrosequencing	28 of 33 (84.8%)	20 of 78 (25.6%)	–
Targeted PCR + MiSeq	52 of 61 (85.2%)	10 of 73 (13.7%)	1 of 12 (8.3%)
Phase 3: Resequencing of 325 mutations to evaluate whether AAFs≤ 40%			
Targeted PCR + MiSeq	159 of 164 (97.0%)	3 of 17 (17.6%)	4 of 144 (2.8%)

Rates at which predicted PZMs from WES were also found to be *de novo* with unequal AAFs using three different technologies.

The Pearson's correlations between AAFs detected from WES and those detected by the three resequencing technologies ranged from 0.52 to 0.58, apparently mainly reflecting the relatively low coverage of and hence imprecise AAFs from WES. In contrast, AAFs determined using CloneSeq and targeted PCR with MiSeq were more highly correlated with one another, at 0.85 (Fig. 1c). Although CloneSeq is an excel-lent standard for measuring the AAF of PZMs, it is low-through put and expensive. Our data suggest that targeted PCR with MiSeq is an acceptable alternative that is higher throughput. AAFs determined with Pyroseq showed lower correlation with CloneSeq, at 0.63. In particular, Pyroseq did not correlate well with CloneSeq at lower AAFs (Fig. 1c), for example, AAFs ≤ 40% (Pearson's correlation = 0.64), unlike targeted PCR with MiSeq (Pearson's correlation = 0.92), suggesting a larger variation in detecting lower AAFs using Pyroseq.

We also tested 82 de novo mutations using Sanger sequencing and confirmed 73 of them (or 89%) as genuine de novo mutations, i.e., the mutations were not present at a detectable AAF in the parents' DNA samples. We reconfirmed this initial result using targeted PCR with MiSeq for another 327 de novo mutations and found that 84.1% of the PZMs from Group C were confirmed to arise de novo. Taken together, our data suggest that approximately 7.5% (= 9.7% (the proportion of de novo mutations detected from WES that were high-confidence PZMs in Group C) × 0.84 (the average fraction of genuine de novo mutations in Group C) × 0.92 (the average fraction of genuine PZMs)) of

all detected de novo mutations are likely to be true PZMs detect-able by WES, although the recovery of PZMs would be expected to be higher if the exomes had been sequenced at higher coverage.

It is possible that some potential PZMs might be falsely called as a result of CNVs spanning the region. As such, we performed TaqMan copy number assays on 36 PZMs in Group C to evaluate the rate of PZMs co-occurring with CNVs but did not detect any (Supplementary Table 5), suggesting that the rate at which CNVs might overlap with PZMs is likely to be less than 3%.

PZMs were frequently missed with previous pipelines

Despite the lower overall rate of called de novo mutations using our approach compared to previous studies, we found that most PZMs in Group B had not been previously identified (617 of 1,113 PZMs or 55.4%; Fig. 1d) and an even higher proportion of PZMs in Group C had not been previously reported (390 of 468 PZMs or 83.3%). This suggests that the previous pipelines were more likely to detect gDNMs found only in Group A and that our approach detects with high specificity many PZMs not previously identified, presumably because these PZMs might have been marked as variants with lower quality and were more likely to be flagged as falsely called variants, despite being readily confirmed by complementary technologies. Our data indicate that over 84.8% of the high-confidence PZMs in Group C were confirmed to be bona fide PZMs through the resequencing experiments and that 83.3% of the high-confidence PZMs were not previously reported.

PZMs differ from gDNMs and cancer somatic mutations

Analysis of the mutational properties of PZMs revealed several features that differentiate them from gDNMs. PZMs are enriched on the anti-sense strand (relative to transcription) compared to gDNMs (OR = 1.30, 95% confidence interval (CI) = [1.07, 1.58] for Group C; Supplementary Table 6). Antisense strand bias typically reflects the inherent bias of transcription-coupled nucleotide excision repair, which has a higher fidelity on the sense strand. This results in a higher accumulation of mutations on the antisense strand11, and it is likely that PZMs arise at least in part from this mechanism, similarly to the formation of somatic mosaic mutations in cancers described in previous reports12.

The most common types of mutations among gDNMs and PZMs are C-T and G-A mutations. It has been reported that there is a strong preference for mutations from A to C or T to G in the nucleosome core13, and we observed a similar enrichment of A-C and T-G mutations in PZMs compared to gDNMs (OR = 2.23, 95% CI = [1.64, 2.99] for Group C; Supplementary Table 7). In particular, we found that the enrichment of A-C mutations was predominantly on the sense strand, whereas the enrichment of T-G mutations was predominantly on the antisense strand (Supplementary Table 8). This is a distinct mutational profile from those reported for somatic mosaic mutations in cancers12, but it is suggestive that the enrichment of such mutations in the nucleosome core might affect chromatin remodeling, a process that has been previously found to be perturbed in ASD14.

Somatic mutations discovered in cancers have also been reported to be associated with late DNA replication12. We correlated PZMs against DNA replication timing during the S phase15 and compared these against the gDNMs found only in Group A (Supplementary Table 9). We observed a similar trend for PZMs with late replication timing (OR = 1.36, 95% CI = [0.83, 2.14] for Group C) but not for those with early replication timing (OR = 0.88, 95% CI = [0.72, 1.07] for Group C). However, the association of these PZMs with late replication timing was substantially less than that reported in cancers16 and was not statistically significant (Fisher's exact test, P = 0.2 for Group C;

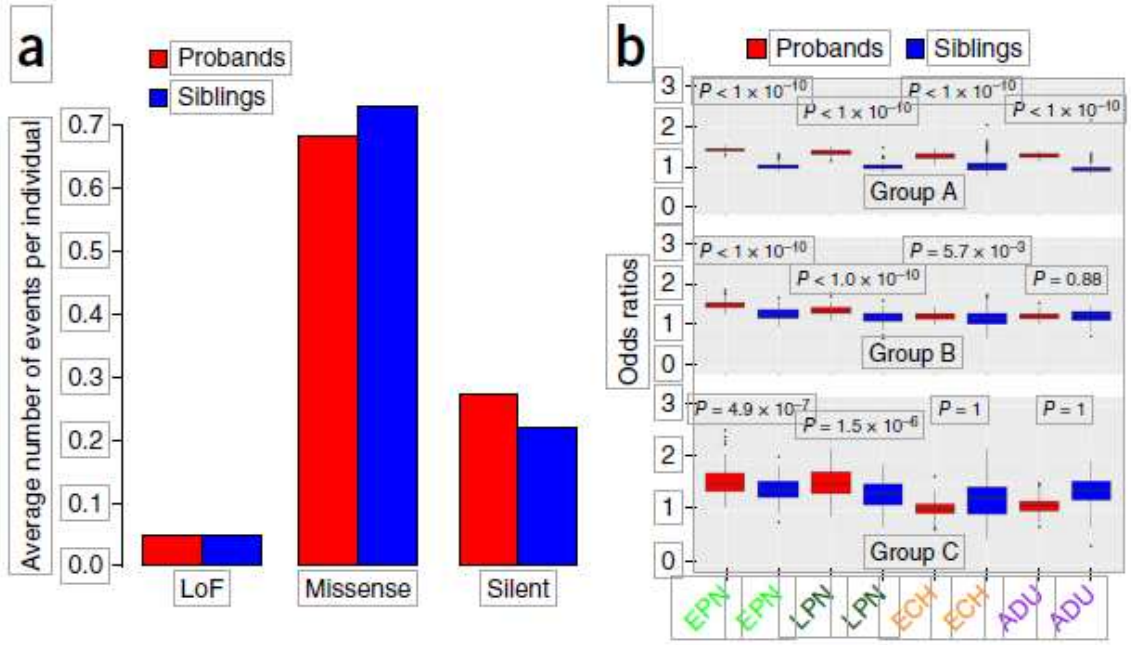

Figure 2 Postzygotic mutations in ASD show excess deleterious mutations in critical exons of genes expressed during early brain development.

(a)There is no statistically significant global excess of Group C PZMs in the probands (red) compared to their unaffected siblings (blue; hypergeometric P = 0.32 for fraction of LoF variants in probands compared to siblings).

(b)As expected, there are highly significant excesses in overall gDNMs(Group A) for genes expressed in prenatal and adult brains. For Groups B and C, representing potential and high-confidence PZMs, there is a strong excess of LoF and missense mutations in critical exons that are expressed in early prenatal (EPN) and late prenatal (LPN) brains (one-tailed Wilcoxon rank-sum test, P < 1 × 10−5) but not in early childhood (ECH) or adult (ADU) postmortem brain samples in the probands (one-tailed Wilcoxon rank-sum test, P > 1 × 10−5). The horizontal lines in the boxplots indicate medians; the box limits indicate first and third quantiles; and the vertical whisker lines indicate minimum and maximum values.

Supplementary Table 9). Together, these results highlight some unique features of the PZMs. Our data suggest that the mechanisms generating PZMs and their mutational profile are distinct from those of gDNMs. Also, while PZMs detected in blood and somatic mosaic mutations in cancers accumulate preferentially on the antisense strand, they differ in their preferences for nucleotide base substitutions.

It has been previously reported that gDNMs are enriched on the paternal haplotype, and similarly, we observed a 1.69-fold excess of mutations in Group A on the paternal haplotype (1,321 paternal versus 781 maternal, binomial P = 1.50 × 10−32; Supplementary Table 10). In contrast, the high-confidence mosaic mutations in Group C did not show any significant excess of mutations on the paternal compared to maternal haplotypes (90 paternal versus 78 maternal, 1.15-fold, binomial P = 0.2). This confirmed that the mutations detected in Group C were likely to be enriched for true PZMs compared to the larger set of Group A mutations.

An excess of deleterious PZMs is found in brain-expressed critical exons in ASD probands

We next investigated whether PZMs might contribute to ASD risk. We first analyzed all de novo LoF mutations in Group A and found the expected excess in probands compared to unaffected siblings, similarly to previous reports4,5. However, the LoF PZMs from Groups B and C did not show an excess in probands versus siblings (Fig. 2a and Supplementary Table 11). When comparing missense PZMs predicted to be deleterious using three in silico tools (PolyPhen2 (ref. 17), SIFT18 and CADD19), we found more de novo missense mutations predicted to be deleterious in Groups A and B in probands compared to siblings but no enrichment for PZMs in Group C (hyper-geometric P = 0.024 for Group A, P = 0.041 for Group B and P = 0.32 for Group C; Supplementary Table 12).

We wondered whether PZMs might contribute to ASD risk by selectively affecting genes expressed in the brain that are subjected to strong purifying selection. It has been previously shown that analysis

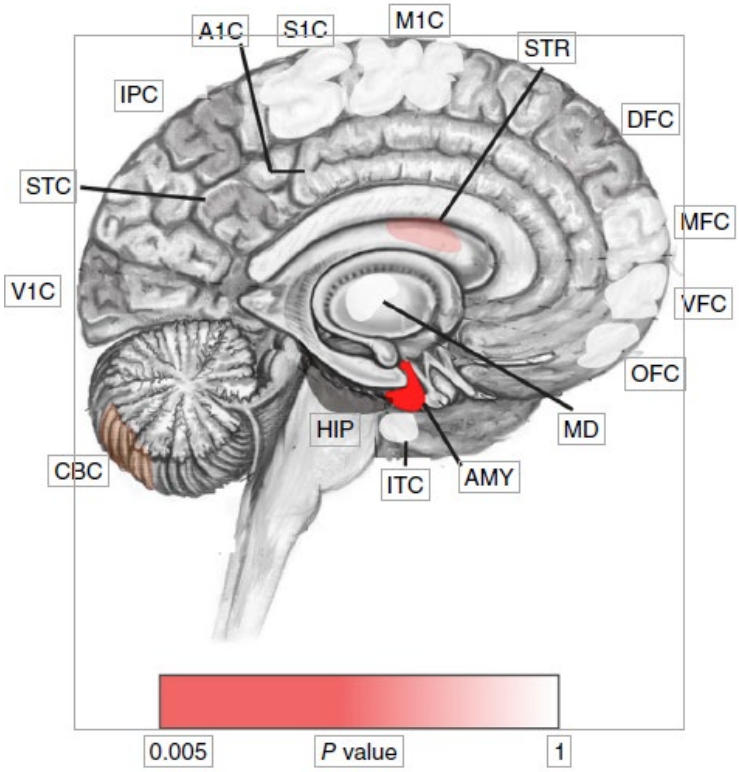

Figure 3 Postzygotic mutations implicate the prenatal amygdala in ASD. The figure shows the spatial representation of regions enriched for genes carrying PZMs found in Group C from the probands. The Group C PZMs in the probands showed the greatest enrichment for expression in the amygdala (one-tailed Wilcoxon rank-sum test, P = 5.4 × 10−3). V1C, primary visual cortex; STC, posterior (caudal) superior temporal cortex; IPC, posterior inferior parietal cortex; A1C, primary auditory cortex; S1C, primary somatosensory cortex; M1C, primary motor cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; VFC, ventrolateral prefrontal cortex; OFC, orbital frontal cortex; ITC, inferolateral temporal cortex; AMY, amygdaloid complex; CBC, cerebellar cortex; HIP, hippocampus; MD, mediodorsal nucleus of thalamus; STR, striatum. of 'critical exons'—i.e., those that are depleted for deleterious mutations in normal individuals—permits higher sensitivity in detecting differences in gDNMs and shows an excess of deleterious PZMs in probands versus unaffected siblings in critical exons expressed in the brain20. In line with previous evidence, we observed an enrichment of LoF and missense mutations from Groups A and B found in critical exons in probands versus unaffected siblings (Fig. 2b). Notably, we also observed an enrichment of high-

confidence LoF and missense PZMs from Group C in the probands compared to siblings, further supporting the association of some of these PZMs with ASD.

Mutations in Group A that fell within critical exons were enriched in probands compared to their unaffected siblings in genes expressed across all developmental epochs: early (≤ 16 weeks after conception) and late prenatal brains (>16 weeks after conception), early childhood (<15 years) and adulthood (≥ 15 years). Mutations in Groups B and C that fell within critical exons were enriched in probands for genes expressed in prenatal and early childhood brains but not in adult brains (Fig. 2b), suggesting a particular enrichment for these genes in processes that occur prenatally, including neurogenesis, neuronal migration, dendritogenesis and synaptogenesis. Assessment of PZMs in Group C that fell in critical exons across 16 brain regions during prenatal development pinpointed the amygdala as the top brain region where PZMs in critical exons were enriched in probands compared to unaffected sib-lings (Wilcoxon rank-sum test; P = 5.4 × 10−3; Fig. 3 and Table 2). Our data suggest that further analyses of PZMs in ASD may begin to unveil brain regions important for the pathophysiology of the disorder.

An excess of recurrent PZMs in genes found in probands implicate these genes in ASD

In probands, 27 of 735 genes (3.7%) showed recurrent nonsynonymous PZMs, versus 2 of 322 genes (0.62%) with recurrent nonsynonymous

Table 2 Regions enriched for PZMs in Group C in probands vs. their unaffected siblings

Brain Region	Group C Wilcoxon test *P*
Amygdala (AMY)	5.4×10^{-3}
Striatum (STR)	0.065
Cerebellar cortex (CBC)	0.093
Hippocampus (HIP)	0.10
Posteroinferior parietal cortex (IPC)	0.27
Primary visual cortex (V1C)	0.43
Primary auditory cortex (A1C)	0.48
Primary motor cortex (M1C)	0.49
Mediodorsal nucleus of thalamus (MD)	0.59
Posterior superior temporal cortex (STC)	0.69
Medial prefrontal cortex (MFC)	0.71
Ventrolateral prefrontal cortex (VFC)	0.71
Inferior temporal cortex (ITC)	0.80
Dorsolateral prefrontal cortex (DFC)	0.96
Orbital prefrontal cortex (OFC)	0.96
Primary somatosensory cortex (S1C)	1

The P values reported are calculated using a two-tailed Wilcoxon rank-sum test. PZMs in siblings, representing a 6.1-fold excess of genes with recur-rent nonsynonymous PZMs in the probands (95% CI = [1.52,53.2], Fisher's exact test, P = 0.0035; permutation, P = 0.0037). This strongly suggests that some of these genes with recurrent nonsynonymous PZMs are relevant for ASD risk.

Given our finding that some genes with recurrent nonsynonymous PZMs were likely to confer risk for ASD, we focused on these genes containing recurrent nonsynonymous PZMs. We obtained a back-ground set of 84,448 variants that were privately inherited (i.e., variants that were not found in our controls, consisting of parents and siblings, or in control databases such as the Exome Variant Server but that were inherited from a parent in an affected or unaffected offspring; Online Methods and Supplementary Table 13). Amongst these, we selected a subset with AAF ≤ 80% from the expected modal AAF to obtain a background rate of PZMs in each gene. In addition, we filtered our

genes in regions with segmental duplications as described previously10, allowing us to exclude genes with falsely called PZMs due to segmental duplications or common CNVs.

We found 27 genes with recurrent nonsynonymous PZMs in the probands, of which two genes (KLF16 and MSANTD2) harbored more PZMs than expected genome-wide based on their background rates (hypergeometric P < 0.05/18,782 = 2.7 × 10−6; Table 3). Among the 27 found in ASD probands4,5, and de novo mutations in HNRNPU have been associated with epileptic encephalopathies21. Our approach detects genes with more recurrent, nonsynonymous PZMs than expected from the number of falsely called mutations. There are several reasons why this might occur; for instance, some genes might be less likely to be repaired and thus may tend to accumulate PZMs. Nonetheless, multiple PZMs within well-documented neurodevelopmental disease genes like SCN2A and HNRNPU provide strong evidence that at least some of the postzygotic mosaic mutations can predispose individuals to ASD.

Among the top genes with recurrent nonsynonymous PZMs in probands, eight of ten were expressed in the brain (Supplementary Table 14), whereas two of the bottom ten genes with recurrent nonsynonymous PZMs in probands showed brain expression. Although we found two genes with recurrent nonsynonymous PZMs in unaffected siblings, neither of these genes was significant genome-wide (top hypergeometric P = 8.5 × 10−5; Supplementary Table 15). gDNMs in ASD probands have been reported in genes that are less tolerant of mutation, defined by lower residual variation intolerance scores22.

Table 3 List of top ten genes with recurrent nonsynonymous PZMs from Group B

	Expected	Observed	Hypergeometric *P*
KLF16	**0 of 84,448**	**2 of 571**	**$<1 \times 10^{-6}$**
MSANTD2	**1 of 84,448**	**2 of 571**	**$<1 \times 10^{-6}$**
POLA2	**2 of 84,448**	**2 of 571**	**4.6×10^{-5}**
SMARCA4	**11 of 84,448**	**3 of 572**	**4.9×10^{-5}**
AZGP1	**4 of 84,448**	**2 of 571**	**2.7×10^{-4}**
CNGB3	5 of 84,448	2 of 571	4.5×10^{-4}
HNRNPU	**5 of 84,448**	**2 of 571**	**4.5×10^{-4}**
SCN2A	**5 of 84,448**	**2 of 571**	**4.5×10^{-4}**
EPPK1	58 of 84,448	4 of 571	6.6×10^{-4}
CARD11	**7 of 84,448**	**2 of 571**	**9.4×10^{-4}**

Genes with recurrent nonsynonymous PZMs from Group B found in the probands (observed), with the observed number of mosaics that are inherited (expected), as well as the hypergeometric test *P* value. The genes that are expressed in the brain are bolded.

We found that genes with recurrent nonsynonymous PZMs in probands that scored highest, i.e., those that had the lowest hypergeometric P values, showed low residual variation intolerance scores, that is, they were more intolerant of variation (Supplementary Fig. 1). These data all further support a role for some of these PZMs in ASD risk.

It has been repeatedly reported that genes implicated in ASD based on de novo mutations are enriched for targets of the Fragile X mental retardation protein (FMRP)5. We replicated this observation for de novo mutations in Group A (OR = 2.72, CI = [2.35, 3.13], P < 1 × 1−10). We also found a significant enrichment for PZMs in Groups B and C for FMRP target genes (OR = 2.65, CI = [2.04, 3.41], P < 1 × 1−10 and OR = 2.06, CI = [1.30, 3.12], P = 7.7 × 10−4, respectively in groups B and C).

PZMs in SMARCA4 downregulates GRIN2B

One of the genes with recurrent nonsynonymous PZMs is SMARCA4, which encodes BRG1, a critical component of the SWI–SNF chromatin- remodeling complex, which regulates gene expression23. Germline and somatic LoF mutations in this gene have been implicated in a variety of cancers, including rhabdoid tumors and small cell carcinoma of the ovary of hypercalcemic type23 (Fig. 4). On the other hand, germline heterozygous missense mutations in SMARCA4 have been associated with Coffin-Siris syndrome (OMIM #135900), which is characterized by intellectual disability. The absence of LoF mutations in SMARCA4 in Coffin-Siris syndrome suggests that the missense mutations act as gain-of-function or activating mutations, unlike the germline inactivating mutations in cancers24.

We detected and confirmed the three missense mutations in SMARCA4 in the three probands with ASD (p.P143A with AAF 21%, p.I184T with AAF 33% and p.P109L with AAF 36%; Fig. 4a), all predicted to be deleterious using PolyPhen2 and SIFT17,18. The p.P143A mutation had a CADD score of 19.81, while the p.I184T and p.P109L mutations had CADD scores of ≥ 20 (26.4 and 34 respectively). The p.P109L mutation was previously reported as a somatic mutation in a lung carcinoma sample from the COSMIC database (COSM710132)25. CloneSeq on blood-derived DNA for these three individuals with the SMARCA4 mutations confirmed two of the mutations as likely PZMs (p.P143A: 45 alternate out of 164 total colonies, binomial P = 4.9 × 10−9; and p.I184T: 39 alternate out of 118 total colonies, binomial P = 1.5 × 10−4), while the p.P109L mutation is likely germline (p.P109L: 83 alternate out of 164 total colonies, binomial P = 0.59).

All three probands had IQs higher than 70 and were confirmed to not show the typical features of Coffin-Siris syndrome. Whereas most SMARCA4 mutations reported in cancers, such as medulloblastoma, fall within the helicase domains of the protein26, the PZMs in SMARCA4 in ASD probands fell in the N-terminal domain, in a region (between amino acids 1 and 282) that binds CREST27, encoded by SS18L1 (Fig. 4b). The BRG1–CREST complex regulates the NR2B subunit of the ionotropic NMDA glutamate receptor27, encoded by the ASD risk gene GRIN2B4,5.

Therefore, we hypothesized that the PZMs in SMARCA4 might influence the BRG1–CREST interaction and thus the expression of the downstream target GRIN2B. To test the hypothesis, we overexpressed wild-type (WT), p.I184T or p.P143A SMARCA4 in mouse neuroblastoma (N2A) cells and measured the expression of GRIN2B by quantitative PCR. We found that overexpression of either SMARCA4 mutant led to significantly lower expression of GRIN2B compared to WT SMARCA4 (Fig. 4c).

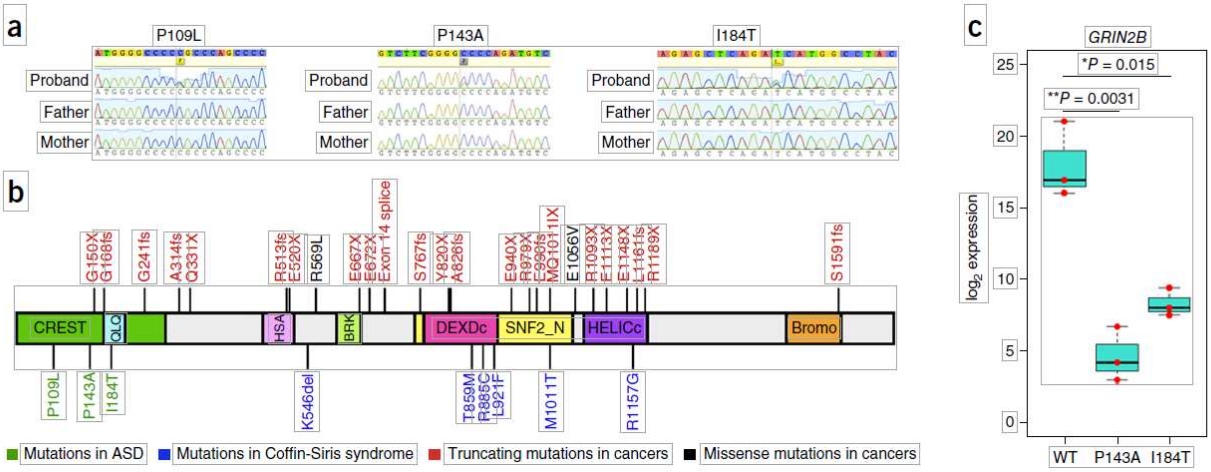

Figure 4 Recurrent nonsynonymous postzygotic mosaic mutations implicate previously uncharacterized genes with more mutations than expected false calls. (a) Sanger sequencing traces for the three SMARCA4 mutations. (b) SMARCA4 mutations reported in cancers, Coffin-Siris syndrome and ASD. (c) qPCR results for GRIN2B after overexpression and selection of WT and mutant (p.P143A and p.I184T) human SMARCA4 in mouse N2A cells, with the values for each replicate experiment (n = 3 each for WT, P143A and I184T) in red dots (unpaired t test, P = 0.0031 for P143A vs. WT;P = 0.015 for I184T vs. WT), showing that these mutations caused abnormal regulation of GRIN2B expression. The horizontal lines in the boxplots indicate medians; the box limits indicate first and third quantiles; and the vertical whisker lines indicate minimum and maximum values.

DISCUSSION

Our systematic analysis of WES from over 5,800 trios found that 7.5% of de novo mutations are PZMs, despite the limited sensitivity of WES to detect PZMs due to its relatively low coverage. We established a pipe-line for detecting and analyzing PZMs, using three independent resequencing technologies, and showed that there was high specificity in our PZM detection. In particular, 84.8–93.3% of the high-confidence Group C PZMs were bona fide PZMs. We also discovered that ASD probands harbor more deleterious PZMs than their unaffected siblings in brain-expressed critical exons, supporting a role for some of these PZMs in ASD risk. Our estimate of 7.5% of de novo mutations being PZMs is similar to the 6.5% rate reported in an earlier cohort of 50 trios with intellectual disability28, as well as to a recently reported estimate of 5.4% in 2,388 families with ASD29. Furthermore, the size of our dataset has allowed us to explore and confirm the role of PZMs in conferring risk to ASD, analyze the mutational characteristics of PZMs and begin to use them to study the spatiotemporal distribution of PZMs in ASD. Our analysis also revealed striking enrichment of PZMs within genes that are clinically relevant to ASD, including the bona fide ASD risk gene SCN2A. The identification of recurrent nonsynonymous PZMs in a small set of genes in ASD probands also provides strong evidence for the clinical importance of PZMs.

The finding that LoF and missense PZMs in critical exons in ASD probands showed enrichment in amygdala expression is intriguing since the amygdala plays key roles in emotional and social responses30, such as conditioned fear. Complete bilateral damage in the amygdala in humans results in impaired social judgement31, reaffirming the importance of the amygdala in regulating social conditioning and learning. An ‘amygdala theory’ of autism32 has been supported by recent work that found impaired neuronal responses in the amygdala in individuals with ASD33. Sexual dimorphism has also been observed in response to testosterone in the amygdala34, which has been proposed to potentially account for some of the gender bias observed in ASD.

We also identified two PZMs in SMARCA4, a gene that encodes a major chromatin factor implicated in cancer and Coffin-Siris syndrome. Both PZMs in SMARCA4 found in the ASD probands fall within the same N-terminal, CREST-binding domain, forming a complex that regulates the activity-dependent expression of key genes implicated in neuronal plasticity27. We discovered that overexpressing SMARCA4 mutants (with p.I184T and p.P143A) reduced the expression of GRIN2B, which encodes a key subunit of the NMDA glutamate receptor that has been previously implicated as an ASD risk gene based on de novo LoFs4,5. This suggests that the PZMs in SMARCA4 might impair the function of glutamatergic synapses35.

It has been reported that de novo CNVs and DNMs associated with ASD are more common in individuals with low nonverbal-IQ scores5. To test the association of IQ with PZM carriers, we analyzed seven probands with recurrent PZMs for which IQ scores were available. Two of these

seven probands with PZMs (or 28.6%) had nonverbal IQs of at least 100, compared to two out of 65 probands (or 3.1%) with recurrent de novo LoF mutations having nonverbal IQs of at least 100, indicating a 9.3-fold excess of probands with higher nonverbal IQs harboring PZMs (hypergeometric test, P = 0.01). This preliminary observation would need replication in a larger number of individuals to test the hypothesis that individuals harboring PZMs might be less severely affected than individuals harboring gDNMs, in terms of cognitive abilities such as IQ, and to test whether PZMs may be overrepresented in a subset of individuals with higher-functioning forms of ASD.

Although the number of probands with IQ data is small, our data suggest that recurrent PZMs were found in individuals with higher IQs than previously reported gDNMs associated with ASD were. This raises the intriguing possibility that some individuals with higher-functioning forms of ASD might harbor PZMs that might be distributed in and affect some but not all regions of the brain, such as the amygdala. This is also consistent with previous observations that higher-functioning individuals, such as unaffected parents, might harbor low levels of parental mosaicism at low AAFs and can transmit these mosaic risk alleles to their affected offspring, which then present as germline mutations in the offspring36. A previous targeted resequencing experiment discovered a mosaic (AAF ~10%) nonsense mutation in the ASD risk gene ADNP in an unaffected parent, providing further anecdotal evidence for this hypothesis37. It is also plausible that some PZMs could create mosaic clinical phenotypes in which the presence of the same mutant allele in the germline would be lethal, such as the AKT1 E17K mutation that causes Proteus syndrome38.

One limitation of our work is that we have not analyzed the potential role of postzygotic CNVs in ASD. Given the strong association of de novo CNVs with ASD2,39,40, it is possible that there might be mosaic CNVs that are involved in ASD, and like the PZMs, mosaic CNVs might be underdetected in previous large-scale genomics studies looking at primarily germline CNVs in ASD. Another area worth pursuing in the future is the role of parental mosaicism in ASD. Such mutations, if present at low AAFs as in the ADNP example, might result in the parents appearing to be clinically unaffected but lead to an increased, recurrent risk for disease in their offspring. It may be useful to survey a large number of unaffected parents (or other control individuals) to understand the rates of mosaicism, and the distribution of AAFs in disease-associated genes, that do not result in a clinical presentation.

Multiple lines of evidence suggest that ASD-associated PZMs detectable in blood samples arise during early development and are enriched in genes expressed in prenatal but not postnatal postmortem brains. Many of the PZMs associated with ASD discovered in blood have relatively high AAFs and are thus likely to have arisen relatively early in development. Our previous studies have shown that functionally neutral PZMs with >5% AAF are likely to be found in multiple tissues8, suggesting that many of the PZMs discovered in blood are likely to be PZMs in brain tissue as well. Given that 83.3% of the high-confidence PZMs were missed using previous algorithms, it may be useful in the future to perform a detailed reanalysis, as well as additional spatiotemporal analyses, on PZMs in other neurodevelopmental and psychiatric disorders, such as intellectual disability, epilepsy and schizophrenia, to understand the role and contribution of PZMs in these disorders.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to all the families who participated in the research, including the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (SSC), the Autism Sequencing Consortium (ASC) and Autism Speaks. We acknowledge the clinicians and organizations that contributed to samples used in this study, including the ASC and SSC principal investigators; the coordinators and staff at the ASC and SSC sites for the recruitment and comprehensive assessment of simplex families; and the ASC, SFARI and NDAR staff for facilitating access to the data sets. This work was supported by a grant from the Simons Foundation (178093, C.A.W.); the National Institutes of Health (NIH) grants R01MH083565, RC2MH089952 and U01MH106883 to C.A.W.; grants R01MH097849, U01MH100233, U01MH100209, U01MH100229, U01MH100239, U01MH111661, U01MH111660, U01MH111658, U01MH111662 and R01MH097849 to the Autism Sequencing Consortium; grants from the Centre for Applied Genomics, the University of Toronto McLaughlin Centre, Genome Canada and Autism Speaks (S.W.S.); Simons Foundation grant (368485, G.M.C.); SRPBS and Brain/MINDS grants from AMED (I.K., B.A., N.O.); grants from the Spanish Ministry of Economy and Competitiveness (M.P.), Instituto de Salud Carlos III (M.P.), PI10/02989 (M.P.), CIBERSAM (M.P.) and ERA-NET NEURON (M.P., C.M.F.), Network of European Funding for Neuroscience Research (M.P.), and Fundación María José Jove and The Institute of Health Carlos III-Fondo de Investigaciones Sanitarias grant project PI13/01136 (A.C.) and the Seaver Foundation. We thank A. Hossain and N. Hatem for their help with sample preparation; F. Zhao and C. Stevens for their help with reprocessing the BAM files; and M. Daly, S. McCarroll, G. Genovese and J. Hirschhorn for comments and suggestions. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. Additional computing support was provided by the Harvard Medical School's Orchestra High-Performance Computing Group, which is partially supported by NIH grant NCRR 1S10RR028832-01. The NHLBI GO Exome Sequencing Project and its ongoing studies produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). C.A.W. is an Investigator of the Howard Hughes Medical Institute; S.W.S. is funded by the GlaxoSmithKline-Canadian Institutes of Heath Research Chair in Genome Sciences at the Hospital for Sick Children and University of Toronto; A.M.D. is supported by the NIGMS (T32GM007753) and NRSA (5T32 GM007226-39); S.D.R. is supported by the Seaver Foundation.

AUTHOR CONTRIBUTIONS

E.T.L. and C.A.W. conceived the project and wrote the manuscript. E.T.L. and M.U. performed the spatiotemporal analyses. E.T.L., S.D.R., Y.C., A.S.K., A.M.D. and S.N.K. performed the resequencing and Sanger sequencing experiments. E.T.L., S.D.R. and X.Z. performed the mutagenesis, overexpression and qPCR experiments. E.T.L. and Y.C. performed the permutations and modeling of background rates. R.S.H., A.P.G. and C.P. performed the data processing and annotation of the variant call files. N.J.M., I.K., B.A., N.O., M.P., C.A., M.J.P., A.C., A.K., C.M.H., L.A.W., A.G.C. and C.M.F. provided additional trio exome sequence data and blood samples for resequencing experiments. E.T.L. and M.F. performed the phasing of mutations. G.M.C., S.W.S., J.D.B. and C.A.W. supervised the project, provided critical comments and edited the manuscript. All authors critically reviewed the manuscript for content.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Gaugler, T. et al. Most genetic risk for autism resides with common variation. Nat. Genet. 46, 881–885 (2014).

2. Sanders, S.J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron 87, 1215–1233 (2015).

3. Pinto, D. et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. 94, 677–694 (2014).

4. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215 (2014).

5. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216–221 (2014).

6. Yu, T.W. et al. Using whole-exome sequencing to identify inherited causes of autism. Neuron 77, 259–273 (2013).

7. Lim, E.T. et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77, 235–242 (2013).

8. Jamuar, S.S. et al. Somatic mutations in cerebral cortical malformations. N. Engl. J. Med. 371, 733–743 (2014).

9. Poduri, A., Evrony, G.D., Cai, X. & Walsh, C.A. Somatic mutation, genomic variation, and neurological disease. Science 341, 1237758 (2013).

10. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487 (2014).

11. Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196 (2010).

12. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518, 360–364 (2015).

13. Prendergast, J.G. & Semple, C.A. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res. 21, 1777–1787 (2011).

14. Cotney, J. et al. The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat. Commun. 6, 6404 (2015).

15. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. 91, 1033–1040 (2012).

16. Woo, Y.H. & Li, W.H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat. Commun. 3, 1004 (2012).

17. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics (eds.Haines, J.L. et al.,) Chapter 7, Unit 7.20 (Wiley, 2013).

18. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081 (2009).

19. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).

20. Uddin, M. et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. Nat. Genet. 46, 742–747 (2014).

21. Carvill, G.L. et al. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat. Genet. 45, 825–830 (2013).

22. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9, e1003709 (2013).

23. Jelinic, P. et al. Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. Nat. Genet. 46, 424–426 (2014).

24. Kosho, T., Okamoto, N. & Coffin-Siris Syndrome International Collaborators Genotype-phenotype correlation of Coffin-Siris syndrome caused by mutations in SMARCB1, SMARCA4, SMARCE1, and ARID1A. Am. J. Med. Genet. C. Semin. Med. Genet. 166C, 262–275 (2014).

25. Forbes, S.A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–D811 (2015).

26. Pugh, T.J. et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature 488, 106–110 (2012).

27. Qiu, Z. & Ghosh, A. A calcium-dependent switch in a CREST-BRG1 complex regulates activity-dependent gene expression. Neuron 60, 775–787 (2008).

28. Acuna-Hidalgo, R. et al. Post-zygotic point mutations are an underrecognized source of de novo genomic variation. Am. J. Hum. Genet. 97, 67–74 (2015).

29. Freed, D. & Pevsner, J. The contribution of mosaic variants to autism spectrum disorder. PLoS Genet. 12, e1006245 (2016).

30. Morris, J.S., Ohman, A. & Dolan, R.J. Conscious and unconscious emotional learning in the human amygdala. Nature 393, 467–470 (1998).

31. Adolphs, R., Tranel, D. & Damasio, A.R. The human amygdala in social judgment. Nature 393, 470–474 (1998).

32. Baron-Cohen, S. et al. The amygdala theory of autism. Neurosci. Biobehav. Rev. 24, 355–364 (2000).

33. Rutishauser, U. et al. Single-neuron correlates of atypical face processing in autism. Neuron 80, 887–899 (2013).

34. Xu, X. et al. Modular genetic control of sexually dimorphic behaviors. Cell 148, 596–607 (2012).

35. Gkogkas, C.G. et al. Autism-related deficits via dysregulated eIF4E-dependent translational control. Nature 493, 371–377 (2013).

36. Campbell, I.M. et al. Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am. J. Hum. Genet. 95, 173–182 (2014).

37. O'Roak, B.J. et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat. Commun. 5, 5595 (2014).

38. Lindhurst, M.J. et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N. Engl. J. Med. 365, 611–619 (2011).

39. Weiss, L.A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. N. Engl. J. Med. 358, 667–675 (2008).

40. Sebat, J. et al. Strong association of de novo copy number mutations with autism. Science 316, 445–449 (2007).

41. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).

42. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92 (2012).

43. Abecasis, G.R. et al. Genomes Project. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).

44. Kang, H.J. et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483–489 (2011).

45. Darnell, J.C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell 146, 247–261 (2011).

46. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. Neuron 74, 285–299 (2012).

47. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. Nature 506, 179–184 (2014).

48. Ramos, P. et al. Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. Nat. Genet. 46, 427–429 (2014).

49. Le Loarer, F. et al. SMARCA4 inactivation defines a group of undifferentiated thoracic malignancies transcriptionally related to BAF-deficient sarcomas. Nat. Genet. 47, 1200–1205 (2015).

50. Tsurusaki, Y. et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat. Genet. 44, 376–378 (2012).

ONLINE METHODS

Standard protocol approval and patient consent. Research performed on samples and data of human origin was conducted according to protocols approved by the institutional review boards of Boston Children's Hospital and Beth Israel Deaconess Medical Center.

Data processing and annotation. The Autism Sequencing Consortium (ASC) performed joint calling of the variants in the 5,947 trios from the ASC and the Simons Simplex Collection (SSC), whose exome sequences have been previously published4,5. The variants were called using two versions of GATK41 (the Unified Genotyper and the Haplotype Caller) and annotated using SnpEff versions 2.0.5 and 3.5 (ref. 42). To remove exomes with inheritance errors, as well as potential artifactual mosaic mutations induced by cell passaging, we removed outlier exomes that had >2 PZMs or >5 de novo mutations from downstream analyses.

PZM detection pipeline applied to the ASC and SSC datasets. We first per-formed joint-calling of the raw files from the previously published and new exomes, in order to obtain standardized data sets for our analyses. Next, we developed a stringent pipeline to call autosomal de novo point mutations from our jointly-called exomes, i.e., mutations that are strictly present in the probands or siblings but not found in both parents. We refer to all de novo mutations as 'Group A', whereas de novo point mutations with AAF equal to or less than 80% of the modal AAF for each cohort are defined as candidate PZMs called 'Group B'. Mutations in Group B where the deviation from the modal AAF was statistically significant (binomial $P \leq 1 \times 10{-}4$) formed 'Group C', the group most likely to be PZMs. The AAF was calculated using: number of alternate reads/(total number of reference + alternate reads).

For our initial analyses, we included all variants that passed a set of quality thresholds (genotype quality, GQ ≥ 20 and alternate read depth ≥ 7). All de novo variants that were observed only once in a proband and were not observed in 6,500 control individuals from the Exome Variant Server (http://evs.gs.washington.edu/EVS/) were included in Group A. In addition, to account for population-specific rare variation, we considered only de novo variants that were not observed in unaffected parents and siblings within each study. Given that there might be differences in capture and sequencing approaches across the various cohorts that could result in overcalling mosaic mutations, we defined PZMs as variants that deviated from the modal AAF (calculated from all de novo variants in Group A) for each cohort, instead of assuming that the modal AAF is 50%. In addition, to reduce false positives as a result of inaccurate realignment, we filtered out PZMs that were within 20 bp of an inherited variant. For the final genes with recurrent nonsynonymous PZMs, we lowered the quality thresholds to alternate DP ≥ 3 in order to screen for additional PZMs that might have been missed and discovered only an additional nonsynonymous PZM in SMARCA4 (I184T with alternate DP = 4).

Resequencing of PZMs. For both the ASC and SSC sequencing projects, DNA derived from mostly blood was used for exome sequencing. We resequenced the PZMs using DNA derived from mostly blood and some lymphoblastoid cells and saliva (from the ASC) or blood and lymphoblastoid cells (from the SSC). For our initial evaluation, we selected 50 de novo mutations for which DNA samples were available (5 from Group A, 28 from Group B and 17 from Group C) and resequenced the mutations using subcloning and Sanger sequencing of the colonies (CloneSeq), targeted PCR followed by MiSeq and pyrosequencing (EpigenDx). Subcloning was performed using the standard protocol with the TA cloning kit (Life Technologies). For targeted PCR, we amplified the genomic

regions around the mutations, performed PCR purification (Qiagen) and sheared the amplicons to ~400-bp fragments before library preparation and sequencing using MiSeq (paired-end, 151 bp).

To obtain an estimate of the rate of de novo mutations detected with our approach, we performed Sanger sequencing for 82 of the PZMs discovered (39 from Group B and 43 from Group C), using samples obtained from the trios and additional family members if available, to confirm the presence of the mutations, as well as the absence of the mutations in the family members, i.e., to confirm the de novo status of the mutations. Given the limitation on detecting low AAFs from Sanger sequencing, we selected variants with AAF ≥ 10% for the Sanger experiments and confirmed 73/82 (89%) as de novo. In particular, 37/39 (94.9%) of the PZMs from Group B were confirmed to arise de novo and 36/43 (83.7%) of the PZMs from Group C were confirmed to arise de novo. In addition, we per-formed targeted PCR with MiSeq for 327 de novo mutations for which parental DNA was available and found that 148/176 (84.1%) of the PZMs from Group C were confirmed to arise de novo, 0/18 (0%) of the PZMs from Group B were confirmed to arise de novo and 131/133 (98.5%) of the gDNMs from Group A were confirmed to arise de novo.

Quantitative PCR for assaying CNVs. For 36 PZMs in Group C for which there are CNVs in the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home) within 2 kb of the PZMs, we selected predesigned primers from ThermoFisher that assay those CNVs. The DNA samples used for these quantitative PCR assays were extracted from whole-blood samples from the Simons Simplex Collection, and quantitative PCR assays were performed by the Biopolymers core facility at Harvard Medical School. The reference assay used was AGO1.

DNA replication timing analyses. We used previously published data15 and mapped the genes from the human genome (hg19 assembly) to the regions with the reported replication timing. We defined early-replicating genes as genes that fall within regions with replication timing $Z \geq 1$ and late-replicating genes as genes that fall within regions with replication timing $Z \leq -1$.

Phasing of de novo mutations. We ran the ReadBackedPhasing tool in GATK to phase the de novo mutations using a 100-kb window around the mutation of interest. Of the 4,846 de novo mutations in Group A, we phased 2,102 (43.4%). Of the 1,113 mutations in Group B, we phased 464 (41.7%) and of the 468 mutations in Group C, we were able to phase 168 (35.9%).

In silico prediction for missense mutations. We used three different tools (PolyPhen217, SIFT18 and CADD19) to obtain in silico predictions for the mis-sense mutations. We defined ‘deleterious mutations’ as all mutations that were predicted by PolyPhen2 to be ‘probably damaging’, by SIFT to be ‘damaging’ and had CADD scores of ≥ 20. We further defined ‘benign mutations’ as all mutations that were predicted by PolyPhen2 to be ‘possibly damaging’ or ‘benign’, by SIFT to be ‘tolerated’ and had CADD scores of < 20.

Critical exon analyses. We used whole-genome sequencing data from the 1,000 Genomes Project43 to compute the burden of rare missense and loss-of-function mutations for each exon. Furthermore, exon level expression data from RNA sequencing was obtained for 524 brain tissues (prenatal and postnatal postmor-tem donors) from the BrainSpan project44. To classify critical exons, we computed the 75th percentile of brain expression and mutational burden for each exon, as described in Uddin et al.20. In short, a critical exon is defined as an exon where expression is high (>75th percentile) and the accumulation of deleterious mutation is low (<75th percentile). For each group (A, B and C) of mutations in the probands and siblings, we first computed the fraction of critical exons with nonsynonymous and synonymous mutations for each brain tissue sample. Next, we computed the odds ratio for each tissue sample by normalizing the fraction of critical exons

detected with the nonsynonymous mutations by the fraction of critical exons detected with the synonymous mutations. Each data point corresponded to a ratio for each expression sample that was inferred from the nonsynonymous/synonymous mutation counts in critical exons.

Inherited variant analyses. To obtain a background rate for comparing the nonsynonymous postzygotic mutations detected from the exomes beyond false calls, we obtained all the inherited variants that were ≤ 1% in the population and selected all variants with AAFs≤ 80% of the modal AAF calculated from the de novo mutations. We used these inherited variants that deviated from the expected modal AAF for modeling the background rates of obtaining PZMs in each gene, to account for technical biases resulting from amplification, exome capture or sequencing. To evaluate the significance of observing recurrent PZMs in each gene beyond expected false calls, we calculated the hypergeometric test P value by comparing the observed number of PZMs for each gene with the expected gene-specific background mutation rates (Supplementary Table 13). The genome-wide threshold was calculated as $P < 0.05/18,782 = 2.7 \times 10{-}6$ as there are 18,782 annotated genes in the data.

Spatial and temporal analyses. To evaluate the distributions of mutations found in genes that are expressed in postmortem brains (prenatal and postnatal), as well as in specific regions of the brain, we downloaded RNA sequencing data from the BrainSpan project44 (http://www.brainspan.org). For the spatial analyses, we focused on 16 brain regions (V1C, primary visual cortex; STC, posterior (caudal) superior temporal cortex; IPC, posterior inferior parietal cortex; A1C, primary auditory cortex; S1C, primary somatosensory cortex; M1C, primary motor cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; VFC, ventrolateral prefrontal cortex; OFC, orbital frontal cortex; ITC, inferolateral temporal cortex; AMY, amygdaloid complex; CBC, cerebellar cortex; HIP, hippocampus; MD, mediodorsal nucleus of thalamus; and STR, striatum).

FMRP target data set. To evaluate the enrichment of FMRP targets, we obtained a list of the transcripts published in Darnell et al.45 that were previously used to evaluate de novo variants in ASD46 and schizophrenia47.

Residual variation intolerance score (RVIS) analyses. We downloaded the RVIS gene scores based on variants reported in the ExAC database with allele frequencies up to 1% (http://genic-intolerance.org/data/RVIS_Unpublished_ ExAC_May2015.txt; accessed 11 October 2016).

Permutations for comparing proband PZMs to sibling PZMs. Group B contained 786 PZMs B found in probands, resulting in 27/735 genes with recurrent nonsynonymous PZMs. Conversely, Group B also contained 327 PZMs found in unaffected siblings, resulting in 2/322 (0.62%) genes with recurrent nonsynonymous PZMs. To evaluate the significance of the excess of recurrent PZMs found in probands compared to recurrent PZMs found in siblings, we randomly sampled 327 PZMs from the 786 PZMs discovered in the probands. We ran100,000 permutations, 367 of which resulted in the proportion of recurrent genes being less than or equal to 2/322.

Mutations in SMARCA4. We compiled a subset of germline and somatic mutations that were reported in cancers48,49, as well as in CoffinSiris syndrome50.

Mutagenesis of SMARCA4 plasmid. We used the human SMARCA4 transcript variant 3, cloned into a pCMV6-AC-GFP backbone (Origene cat no. RG219258). The primers used for the mutagenesis were designed using the Agilent QuikChange design tool, and mutagenesis was performed using the standard protocol with the Agilent QuikChange II XL kit. All mutants were confirmed using Sanger sequencing and plasmids were extracted using the endotoxin-free QIAGEN Plasmid Maxi Kit. We attempted mutagenesis for all three SMARCA4 mutants (c.326C>T or p.P109L, c.427C>G or p.P143A

and c.551T>C or p.I184T), but only two of the mutagenesis experiments resulted in colonies (c.427C>G and c.551T>C). We repeated the mutagenesis experiments for the c.326C>T mutant using the Q5 Site-Directed Mutagenesis Kit (New England BioLabs), which again did not results in any colonies.

The primers used for the p.P143A (c.427C>G) QuikChange mutagenesis experiment are:

Forward: 5'- GAAGACATCTGGGCCCCCGAAGACGGG -3'; and Reverse: 5'- CCCGTCTTCGGGGGCCCAGATGTCTTC -3'.

The primers used for the p.I184T (c.551T>C) QuikChange mutagenesis experiment are:

Forward: 5'- CATCTTGTAGGCCATGGTCTGAGCTCTGAGCTG -3'; and Reverse: 5'- CAGCTCAGAGCTCAGACCATGGCCTACAAGATG -3'.

Overexpression of SMARCA4 plasmids in N2A cells. P4 mouse neuroblastoma (N2A) cells commercially available from ATCC were tested negative for myco-plasma and passaged in DMEM with L-glutamine, 4.5g/L glucose and sodium pyruvate (Thermo Fisher Scientific) with 10% fetal bovine serum (Thermo Fisher Scientific) and 1% Penicillin Streptomycin (Thermo Fisher Scientific). We transfected 24 µg of WT or mutant plasmids into 90% confluent N2A cells in 10-cm tissue culture plates using Lipofectamine 2000 (Life Technologies). The transfections for each plasmid (WT and two mutants) were performed in triplicates. Selection was performed by adding 1,000 µg/mL of G418 antibiotic (Life Technologies) 24 h after transfection to each plate for 10 d, exchanging fresh antibiotics every 3 d. Three additional plates of WT N2A cells were grown without selection as controls.

RNA extraction and qPCR. The N2A cells were dissociated using 0.05% Trypsin-EDTA (Life Technologies) and washed with PBS (Life Technologies). RNA extraction was performed using an Ambion PureLink RNA Mini Kit (Life Technologies) and cDNA synthesis was performed using a SuperScript III First-Strand kit (Life Technologies). KAPA SYBR FAST qPCR master mix was added to 1 µg of cDNA and 1 µL per 10 µM each of forward and reverse primers for the qPCR experiments.

The primers used for the mouse ACTB qPCR experiment were:

Forward: 5'- GGCTGTATTCCCCTCCATCG -3' and Reverse: 5'- CCAGTTGGTAACAATGCCATGT -3'.

The primers used for the mouse GRIN2B qPCR experiment were:

Forward: 5'- CAGCAAAGCTCGTTCCCAAAA -3' and Reverse: 5'- GTCAGTCTCGTTCATGGCTAC -3'.

To obtain the log2 expression levels for GRIN2B, we first calculated the ΔCt = CtGRIN2B – CtACTB for all the qPCR results obtained from mutants, WT and controls, and then normalized the log2 expression levels by calculating ΔΔCt = ΔCtcontrol – ΔCt(mutant or WT).

Statistics. To compare the strand bias among mutations, we calculated the two-tailed Fisher's exact test P values for the numbers of mutations found on the sense and antisense strands in Groups B and C compared to the numbers of mutations in Group A (exact numbers are shown in Supplementary Table 6).

To compare the differences in mutational properties, we calculated the two-tailed Fisher's exact test P values for the numbers of A>C and T > G mutations in Groups B and C to the numbers of mutations in Group A (exact numbers are shown in Supplementary Table 7).

To compare the strand-specific differences in mutational properties, we calculated the two-tailed Fisher's exact test P values for the numbers of A>C and T > G mutations found on the sense and antisense strands in Groups B and C to the numbers of mutations in Group A (exact numbers are shown in Supplementary Table 8).

To compare the association of PZMs with replication timing, we calculated the two-tailed Fisher's exact test P values for the numbers of mutations with early or late replication times in Groups B and C compared to the numbers of mutations in Group A (exact numbers are shown in Supplementary Table 9).

To compare the enrichment of mutations on the paternal or maternal haplotypes, we calculated the binomial test P values for the numbers of mutations in Groups A–C (exact numbers are shown in Supplementary Table 10).

To compare the functional distribution of mutations in probands versus unaffected siblings, we calculated the hypergeometric P values for the numbers of mutations in Groups A–C for probands and siblings (exact numbers are shown in Supplementary Table 11).

To compare the rates of predicted deleterious missense mutations in probands compared to siblings, we calculated the one-tailed Fisher's exact test P values for Groups A–C (exact numbers are shown in Supplementary Table 12).

To prioritize the genes with recurrent nonsynonymous PZMs found in the probands, we calculated the hypergeometric P values for each gene (exact num-bers are shown in Supplementary Tables 13 and 14). Similarly, we calculated the hypergeometric P values for each gene with recurrent nonsynonymous PZMs found in the unaffected siblings; exact numbers are shown in Supplementary Table 15.

No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications4,5. Data collection and analysis were not performed blind to the conditions of the experiments. A Supplementary Methods Checklist is available.

Code availability. All analyses were performed using custom Perl and R scripts, which are available on reasonable request. The code and scripts have also been uploaded to https://pgpresearch.med.harvard.edu/mosaic/.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

BrainSpan Project: http://www.brainspan.org.

Exome Variant Server: http://evs.gs.washington.edu/EVS/.

RVIS: http://genic-intolerance.org/data/RVIS_Unpublished_ExAC_May2015.txt.

Database of Genomic Variants: http://dgv.tcag.ca/dgv/app/home.

Accession code: dbGAP: raw data from the ASC exome sequence data, phs000298.v3.p2.