nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-020-00765-6

The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing

Rachel E. Rodin[1,2,23], Yanmei Dou[3,23], Minseok Kwon[3], Maxwell A. Sherman[3,4,5,6], Alissa M. D'Gama[1,2], Ryan N. Doan[1], Lariza M. Rento[1,2], Kelly M. Girskis[1,2], Craig L. Bohrson[3], Sonia N. Kim[1,2], Ajay Nadig[1,2], Lovelace J. Luquette[3], Doga C. Gulhan[3], Brain Somatic Mosaicism Network*, Peter J. Park[3,5]✉ and Christopher A. Walsh[1,2,6]✉

We characterize the landscape of somatic mutations—mutations occurring after fertilization—in the human brain using ultra-deep (~250×) whole-genome sequencing of prefrontal cortex from 59 donors with autism spectrum disorder (ASD) and 15 control donors. We observe a mean of 26 somatic single-nucleotide variants per brain present in ≥4% of cells, with enrichment of mutations in coding and putative regulatory regions. Our analysis reveals that the first cell division after fertilization produces ~3.4 mutations, followed by 2–3 mutations in subsequent generations. This suggests that a typical individual possesses ~80 somatic single-nucleotide variants present in ≥2% of cells—comparable to the number of de novo germline mutations per generation—with about half of individuals having at least one potentially function-altering somatic mutation somewhere in the cortex. ASD brains show an excess of somatic mutations in neural enhancer sequences compared with controls, suggesting that mosaic enhancer mutations may contribute to ASD risk.

Somatic mutations, also referred to as mosaic mutations, are acquired postfertilization and are present in a subset of an individual's cells, marking only cells that are descended from the originally mutated cell[1]. Mutations occurring early in development may be present throughout the body, whereas later-occurring mutations are found in progressively smaller subsets of cells. Somatic mutations can be neutral or they can confer a functional advantage or disadvantage to the affected cell. Somatic mutations represent a particularly interesting phenomenon in the brain, as most neurons are postmitotic and will harbor their mutations for the life of the individual. The study of benign clonal somatic mutations in the brain has the potential to reveal insights into mechanisms of normal neurodevelopment. Although function-altering somatic mutations have well-established roles in cancer and some focal epilepsies[2–4], the question of whether mutational mosaicism affects risk of other neuropsychiatric diseases is only beginning to be explored[5–7].

Previous studies have established that single human neurons harbor large numbers of somatic mutations[8]. However, the burden and extent of clonal somatic mutations—somatic mutations present in multiple cells and thus likely to have capacity for functional impact—have until now been difficult to characterize. This is largely due to the fact that most studies of clonal somatic mutation in human brain have been performed in very small sample sizes, based on exomes only[9,10] or with insufficient depth of sequencing. Investigating the genome-wide landscape of clonal somatic mutation within the human brain is crucial to better understand the ways in which developmental mutations shape the adult brain, contribute to normal variation and cause disease.

ASD is a complex and heterogeneous neurodevelopmental disease characterized by impairments in communication and social interactions as well as repetitive behaviors. Several large studies of blood or saliva DNA from ASD families have shown that both de novo germline mutations and exonic somatic single-nucleotide variants (sSNVs) play a role in ASD causation, with anywhere from 5.4% to 22% of mutations that were previously thought to be de novo actually representing postzygotic mutations[11–14]. Due to the scarcity of donated brain tissue, very few studies have investigated somatic mutation in autism brain[12,15]. Deep whole-genome sequencing (WGS) of autism-affected brains has not been available, and therefore the critical question of how noncoding somatic mutations in autism brain might impact disease risk remains unanswered.

Here we present ultra-deep WGS of 59 ASD brains and 15 neurotypical brains, representing the largest such collection ever assembled. An earlier study of ASD brains examined targeted sequencing of 78 genes[15], whereas we perform whole-genome analysis of prefrontal cortex (PFC) from some of the same brains and many additional ones. Our deep WGS data enable detection of sSNVs occurring early in development—mutations that are most likely to alter brain function compared with later-occurring mutations—and accurate estimation of allele fractions, allowing for high-resolution analysis of human brain mutational mosaicism. Our analysis provides insight into the general architecture and

[1]Division of Genetics and Genomics, Manton Center for Orphan Disease Research, and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA. [2]Departments of Pediatrics and Neurology, Harvard Medical School, Boston, MA, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. [5]Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA. [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [23]These authors contributed equally: Rachel E. Rodin, Yanmei Dou. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: peter_park@hms.harvard.edu; Christopher.walsh@childrens.harvard.edu

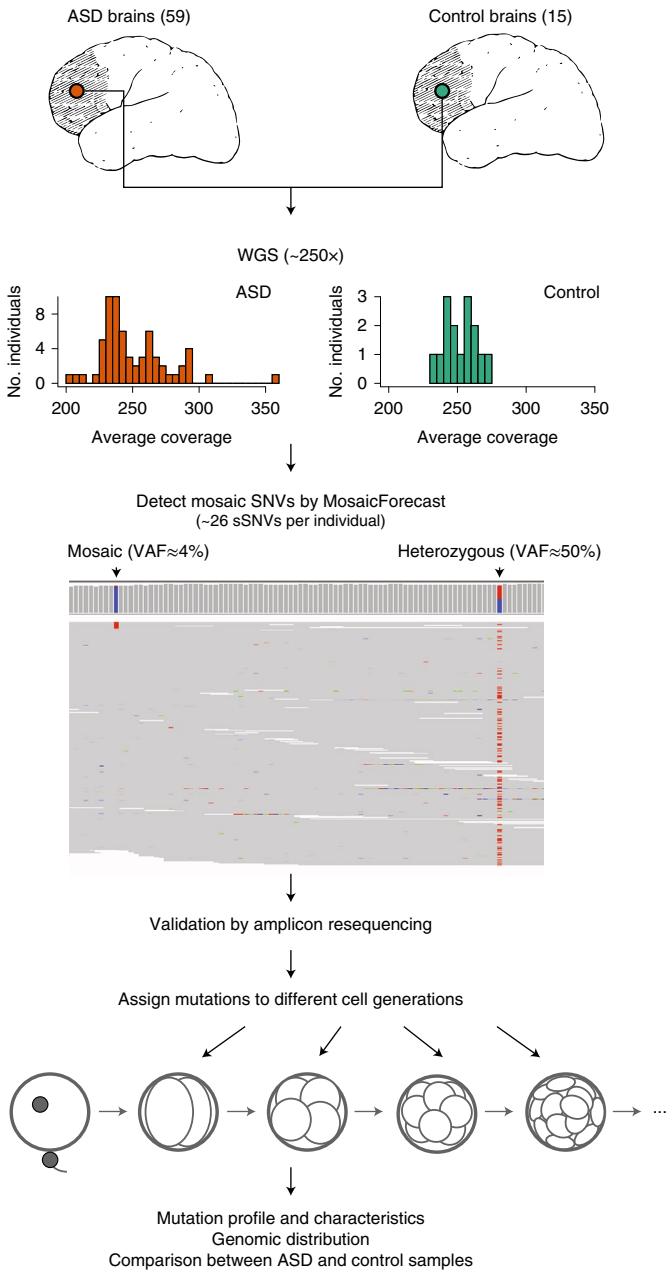

Fig. 1 | Experimental design and genome coverage. DNA was isolated from dorsolateral prefrontal cortex of 59 ASD brains and 15 control brains, then whole-genome sequenced to an average depth of 250×. Germline variants were identified and sSNVs were called using MosaicForecast. A representative set of mutations was validated using targeted amplicon resequencing. Mutations were systematically assigned to cell generations based on VAF.

mutational signatures of somatic mutation in the normal human brain, as well as important implications for genome-wide mosaic mutation in ASD pathogenesis.

Results

Variant discovery and validation. We sampled 59 ASD brains and 15 control brains (Supplementary Table 1), extracting DNA from dorsolateral PFC where available, before performing WGS to an average depth of 250×. The majority of ASD samples and all control samples were sequenced using a PCR-free library preparation, whereas 11 ASD samples were sequenced with a PCR-based preparation. We identified sSNVs and indels in all samples using a machine-learning-based method called MosaicForecast[16], which is optimized to detect somatic mutations in the absence of a matched reference tissue (Fig. 1 and Supplementary Figs. 1 and 2). This is crucial for studying disease samples such as ASD brains, as many donated cases do not include paired nonbrain tissues. We were able to call mutations with variant allele fraction (VAF) down to 2% in PCR-free samples, and to 3% VAF in PCR-based samples; therefore, we had sensitivity to detect variants present in as few as 4–6% of cells in a given section of brain tissue. After stringent filtration, we identified an average of 25.6 ± 10.2 sSNVs per sample (range 10–60; Fig. 2a and Supplementary Table 2). VAFs of detected sSNVs ranged from ~2% to 39%, corresponding to mutation presence in ~4–78% (mean 17%) of cells for autosomal variants (Fig. 2b). Identified sSNV positions were covered at an average of 216×. There was no difference in overall sSNV count between ASD cases and controls.

Deep targeted resequencing of 208 putative sSNVs (Supplementary Table 3), to an average depth of ~50,000× per reaction, demonstrated the accuracy of our mutation-calling algorithm since called sSNVs showed an overall validation rate of 90% (Fig. 2c; 94% for samples sequenced using PCR-free libraries). VAFs from targeted resequencing were highly concordant with VAFs estimated from WGS (Fig. 2d; $R^2 = 0.904$). Analysis of germline heterozygous sites covered in deep resequencing showed over-dispersion of validation VAFs, which was not present in WGS data and was corrected in downstream analyses (Supplementary Fig. 3 and Supplementary Table 4). Since the 11 samples sequenced using PCR-based library preparation had a lower validation rate (Fig. 2c), only sSNVs discovered in PCR-free samples or validated in PCR-based samples were used for downstream analyses. We also validated 86 mosaic indels called with MosaicForecast (Supplementary Table 5 and Supplementary Fig. 4), which were used in regulatory region analyses as described later.

Somatic mutations are enriched in coding and open chromatin regions. At the single-neuron level, it has been shown that sSNVs are enriched in exons, suggesting that transcriptional error may contribute to somatic mutation[17]. However, few studies have examined clonal somatic mutations genome-wide in bulk DNA samples to determine which regions of the genome harbor somatic mutations that arise in early development. We found that across all PCR-free samples, 35 sSNVs were exonic (2.2%), which is about twice as high as expected (Fig. 2e, $P = 0.0003$, two-tailed binomial test, Supplementary Table 6). Among these, 12 were silent, two were protein-truncating and 21 were missense. Approximately 43% (27 of 63) of PCR-free samples had at least one detectable exonic sSNV, with one sample having three exonic mutations (Supplementary Fig. 5) at VAFs detectable in this study. These data suggest that coding regions are particularly vulnerable to somatic mutation during development.

Across all samples, we identified 21 potentially damaging exonic sSNVs, including validated variants in PCR-based samples (Table 1 and Supplementary Table 6). Damaging somatic mutations in mutationally constrained genes (as predicted by the probability of loss-of-function intolerance score (pLI score)) were identified in both cases and controls based on rigorous criteria using 12 different mutation effect prediction tools[18], with the classification of NsynD4 representing the most likely damaging nonsynonymous mutations and the classification of LOF representing predicted loss-of-function mutations (Table 1). Our dataset included two protein-truncating mutations predicted with high confidence to be loss-of-function, although both were in genes predicted to be relatively tolerant to loss-of-function (Supplementary Table 6).

Not surprisingly given our small sample size, the overall burden of exonic sSNVs was similar in our ASD cases and controls, but several of our ASD cases carried damaging mosaic mutations

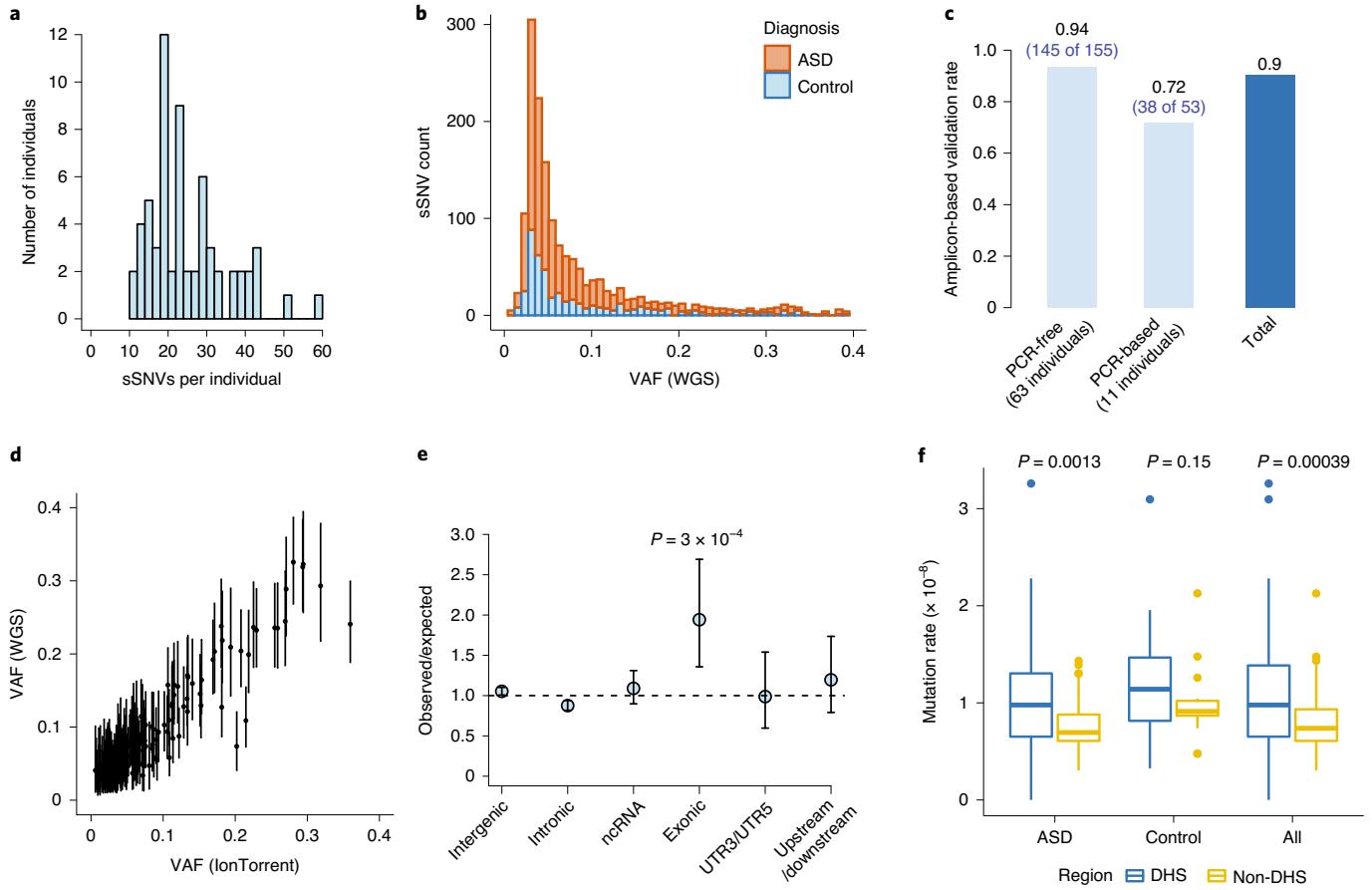

Fig. 2 | Mosaic mutations are present across the genomes of cases and controls. a, Distribution of mosaic mutations per subject. **b**, VAF distribution of all mosaic variants identified in this study (stacked barplot). There are more ASD cases than controls and therefore more total ASD sSNVs, but no difference in allele fraction distribution. **c**, Validation rates in PCR-free and PCR-based samples from deep targeted resequencing of 208 putative mutations. **d**, VAFs from WGS were highly correlated with allele fractions from deep resequencing ($R^2 = 0.904$; $n = 208$ mutations). Error bars indicate 95% CIs calculated with a binomial test. **e**, Number of observed mosaic mutations divided by the expected number of mosaics assuming a uniform mutation rate, among 1,603 PCR-free sSNVs (excluding false positive mutations by deep resequencing validation). Exonic regions show the strongest enrichment for mosaics among several genomic regions (two-tailed binomial test). Error bars indicate 95% CIs calculated with a binomial test. **f**, Noncoding somatic mutations are enriched in DHSs annotated by the Roadmap Epigenomics Project in both the 59 ASD cases and the total dataset (74 cases total; two-tailed Wilcoxon rank sum test). The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle lines correspond to the median values. NcRNA, non-coding RNA.

that may be relevant for the patient phenotype. For instance, a likely damaging missense mutation in *CACNA1A* (c.354 G > T; p.G40W), a gene previously documented to cause autism and intellectual disability in the heterozygous state[19], was found in approximately 10% of brain cells for case UMB1174 (Table 1 and Supplementary Fig. 6). Importantly, *CACNA1A* mutations can also cause epileptic encephalopathy[20], and case UMB1174 was noted to have a seizure disorder in addition to a diagnosis of ASD. While the study size is not well powered to analyze germline ASD mutations, several individual cases nonetheless showed rare, predicted-damaging germline variants in autism risk genes that are likely to contribute to disease, and which may be useful in guiding studies of these samples (Supplementary Table 7). Furthermore, we observed a nominal excess of deleterious germline variants in ASD cases compared with controls when applying a recently published analysis[21] (rate ratio = 4.32, $P = 0.149$, two-tailed Poisson exact test; Supplementary Fig. 7), although as expected from our small sample size, this difference was not statistically significant.

In addition to finding enrichment of sSNVs in exons, we also observed an excess of somatic mutations in areas of open chromatin. We found an increased rate of somatic mutations in noncoding DNase I hypersensitive sites (DHSs) ($P = 0.00039$, two-tailed Wilcoxon rank sum test; Fig. 2f), which often colocalize with regulatory elements such as promoters and enhancers[22,23]. The increased rate of somatic mutation in DHSs could represent heightened vulnerability of regulatory regions to DNA damage and replication errors during early development[24].

Clonal mutation analysis reveals insights into early embryonic development. We examined mutational dynamics in the early embryo by assigning sSNVs to specific cell generations based on their allele fractions (accounting for variable read depth), using a maximum likelihood approach (Methods, Fig. 3a and Supplementary Table 8). We confirmed that our assigned cell generations correlated well with data from three brains that had previously undergone single-cell lineage analysis[8,17] (Fig. 3b). When we estimated mutation rate per cell generation after correcting for detection sensitivity, our data revealed an elevated mutation rate during the first postconception cell division (~3.4 mutations per division), followed by a steady rate of ~2–3 mutations per division in subsequent cell divisions

Table 1 | Damaging mosaic mutations identified in this study

Gene	Mutation (protein)	Case ID	Diagnosis	Library prep method	VAF	Validation result	Pathogenicity prediction	pLI score	Missense Z score	SFARI gene score	Other disease association
Damaging mosaic mutations in constrained genes											
CACNA1A	p.G40W	UMB1174	ASD	PCR-free	0.055	Mosaic	NsynD4	1	7.23	S (syndromic evidence)	Autosomal dominant epileptic encephalopathy
SAFB2	p.R707H	UMB1174	ASD	PCR-free	0.058	Mosaic	NsynD4	0.99	1.32	-	
PTPN12	p.A146V	ABN_JU7U	ASD	PCR-free	0.042	Mosaic	NsynD4	1	−0.05	-	Somatic colon cancer
DCAF8	p.R529W	UMB4899	ASD	PCR-based	0.039	Mosaic	NsynD4	1	3.73	-	Autosomal dominant giant axonal neuropathy
TCERG1	p.D610G	UMB4334	ASD	PCR-based	0.289	Mosaic	NsynD4	1	3.62	-	
COL11A2	p.T1456I	UMB4899	ASD	PCR-based	0.044	Mosaic	NsynD4	1	2.04	-	Deafness
AGAP3	p.P273R	UMB5841	ASD	PCR-free	0.046	N/A	NsynD4	0.99	4.35	-	
FAM13B	p.D561G	UMB5176	ASD	PCR-based	0.047	Mosaic	NsynD4	0.9	−0.53	-	
EIF4G3	p.A1545V	UMB4842	Control	PCR-free	0.044	Mosaic	NsynD4	1	2.3	-	
AEBP2	p.F353I	UMB1712	Control	PCR-free	0.047	Mosaic	NsynD4	0.95	3.74	-	
Damaging mosaic mutations in nonconstrained genes											
DNAH3	p.E2860K	AN09412	ASD	PCR-free	0.127	Mosaic	NsynD1	0	−1.74	3 (suggestive evidence)	
SCARF1	p.P648L	AN09412	ASD	PCR-free	0.058	Mosaic	NsynD2	0	0.63	-	
CAST	p.M1R	M3663M	ASD	PCR-free	0.04	N/A	NsynD2	0	−1.8	-	Autosomal recessive dermatologic disease
COL6A3	p.E330K	UK25363	ASD	PCR-free	0.031	Mosaic	NsynD4	0	−0.3	-	Multiple muscular disease
PARVA	p.K395R	UMB1174	ASD	PCR-free	0.02	N/A	NsynD4	0.34	0.98	-	
GLT8D2	p.T270I	UMB4849	ASD	PCR-free	0.04	Mosaic	NsynD4	0	−0.29	-	
FAM83E	p.R446X	UMB914	Control	PCR-free	0.057	Mosaic	LOF-1[a]	0	−0.38	-	
UNC45A	p.K472N	UMB1024	Control	PCR-free	0.062	Mosaic	NsynD4	0	−0.38	-	
RNF175	p.W102X	UMB1465	Control	PCR-free	0.072	Mosaic	LOF-1[a]	0	−1.07	-	
TTN	p.P5920S	UMB4842	Control	PCR-free	0.158	Mosaic	NsynD4	0	−5.48	3S (suggestive & syndromic evidence)	Cardiomyopathy, muscular dystrophy
PTGDR	p.T69M	UMB5391	Control	PCR-free	0.047	Mosaic	NsynD2	0	1.03	-	

[a]Also predicted high-confidence LOF by LOFTEE. SFARI, Simons Foundation Autism Research Initiative

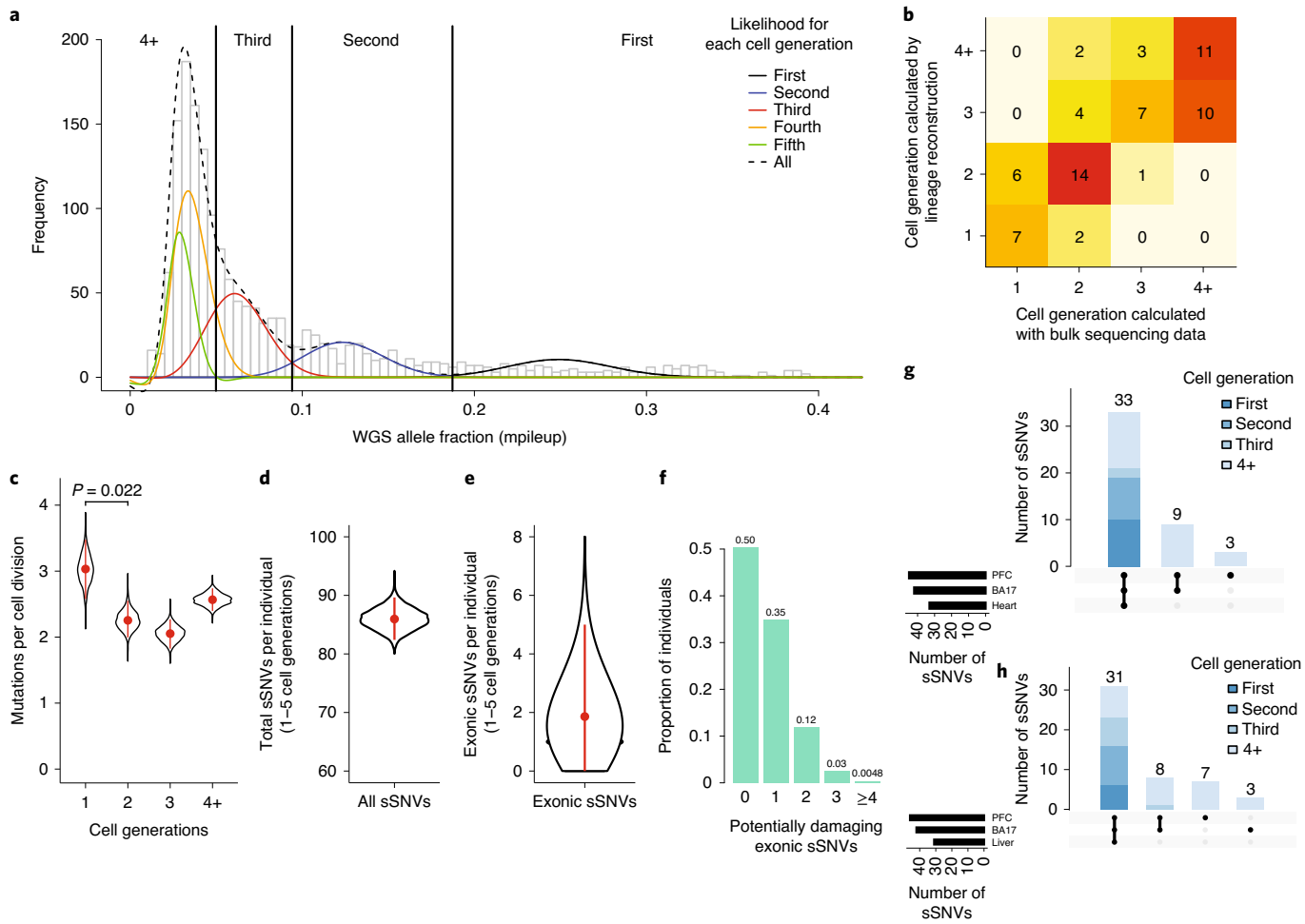

Fig. 3 | Clonal mutation analysis reveals mutational dynamics in the early embryo. a, Clonal somatic mutations map onto a symmetric model of cell division in the early embryo. The black curve represents the likelihood of sSNVs belonging to the first cell generation, blue the second cell generation, red the third cell generation, yellow the fourth cell generation and green the fifth cell generation. **b**, Cell generation assignments for sSNVs were congruent with data from single-cell lineage analysis of three individuals (UMB1465, UMB4643, UMB4638). **c**, Somatic mutations are elevated in the first cell generation of embryogenesis (~3.4 mutations per cell division), then occur at a rate of approximately 2–3 mutations per cell division in subsequent generations. **d**, Based on the mutation rate per cell generation, an average individual would carry ~86 (95% CI: 82–90) sSNVs from the first five cell generations. **e**, Given that ~2.2% of sSNVs in our dataset are exonic, each individual would carry ~1.9 (95% CI: 0–5) exonic sSNVs. Data in **c**–**e** are based on 1,603 PCR-free sSNVs. Plots represent mean and error bars indicate 95% CI calculated with a binomial test. **f**, Assuming ~37% of new exonic mutations are damaging, ~50% of individuals would carry ≥1 damaging exonic mutation from the first five cell divisions, present in roughly ≥2% of cells. **g**, Among mosaic mutations in the occipital lobe and prefrontal cortex for control brain UMB4638, variants assigned to earlier cell divisions are present in wider tissue distributions across the body. **h**, Among mosaic mutations identified in the occipital lobe and prefrontal cortex, variants assigned to earlier cell divisions are also present in wider tissue distributions across the body of control UMB4643.

(Fig. 3c; $P=0.022$ for the difference between the first two cycles based on permutation test; Methods and Supplementary Fig. 8). We note that our validation rate was high across a wide range of VAFs (validation rate = 100% for mosaics with >0.2 VAF; Supplementary Fig. 9), and analysis of single-cell sequencing data[8] confirmed that even high-VAF (>0.2) variants were indeed mosaic and not germline (Supplementary Fig. 10). These findings suggest that the higher mutation rate during the first cell generation is unlikely to be artifactual, and it has been suggested before using similar methods[25].

Based on the positions of the peaks in our mutational VAF distribution, it is possible to infer whether the VAF distribution is more consistent with a symmetric or an asymmetric model of early embryo development. We found that our data slightly favor an asymmetric cell model ($P=3\times10^{-4}$, likelihood ratio test), in which many progenitor cells contribute unevenly to the organism[26] (Supplementary Figs. 11 and 12). Importantly, cell generation assignments changed very little with implementation of the symmetric versus asymmetric models (Supplementary Table 8).

The estimated early embryo mutation rate is largely consistent with other reports, which have estimated roughly 1–3 mutations per cell division in the early embryo[25–27]. Notably, this somatic mutation rate is ~1–5 times higher than the germline mutation rate inferred from de novo germline mutations in maternal haploids, and ~10 times higher than the estimated germline mutation rate in paternal haploids[25,28]. The cell cycle in the early human embryo is relatively short[29], with reduced G1 checkpoint protein expression[30], suggesting one mechanism for the elevated rates of somatic mutation. Furthermore, some mutations in the first cell cycle may represent sites of single-stranded DNA damage in sperm or egg[31,32], repaired to the mutated base during the first cell division, thereby accounting for the elevated mutation rate in the first cell generation after fertilization.

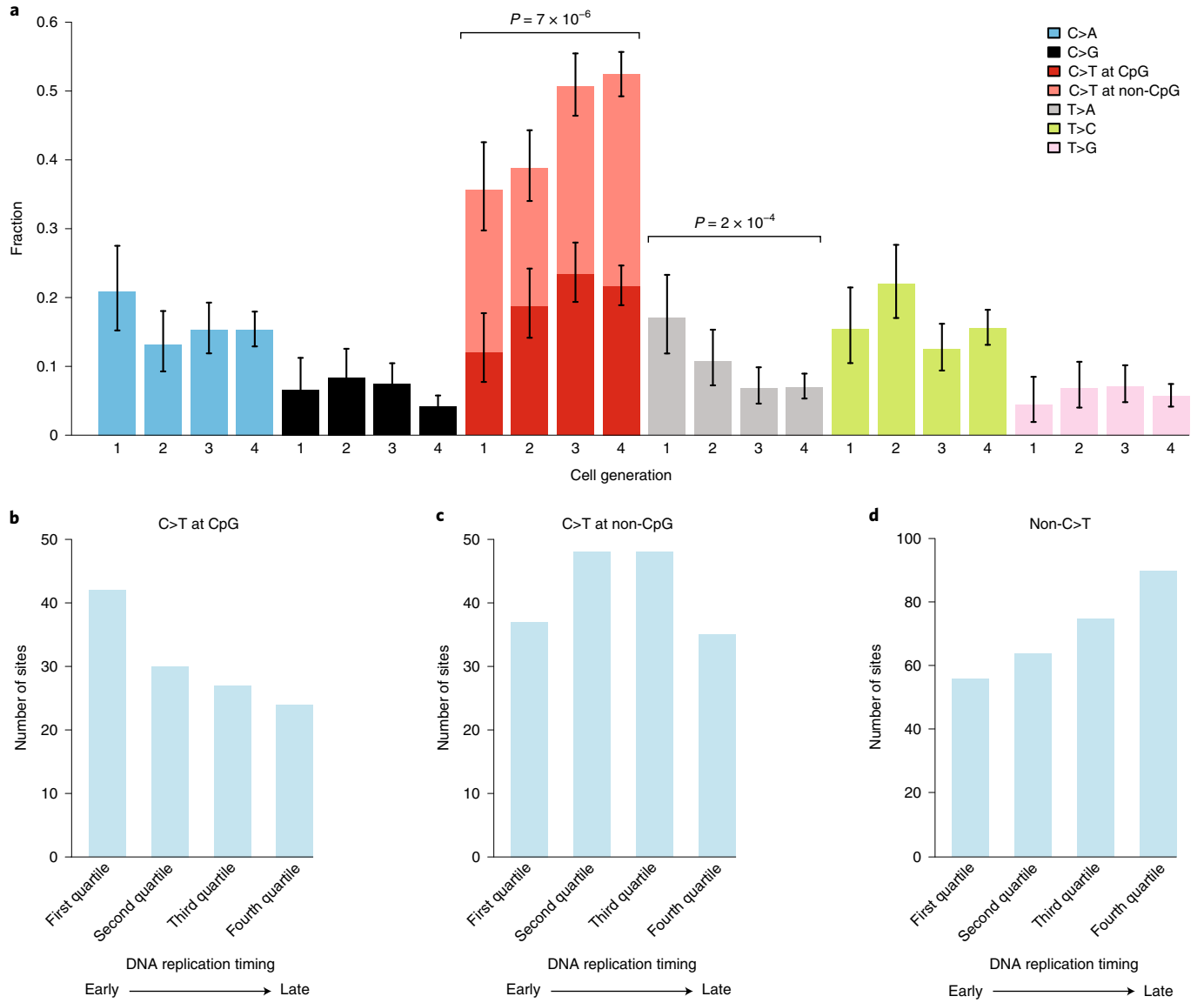

Fig. 4 | Base substitutions vary with cell generation and replication timing. a, There is notable evolution of mutation profiles across the first four cell divisions. Specifically, there is an increase of C > T transitions and decrease of T > A mutations (two-tailed Fisher's exact test). Data include 1,641 mutations (all PCR-free sSNVs with false positives removed and validated PCR-based sSNVs). Error bars indicate 95% CIs calculated with a binomial test. **b**, C > T mutations in CpG dinucleotides tend to occur in earlier-replicating genomic regions. **c**, C > T mutations in non-CpG contexts show no replication timing bias. **d**, All other substitution types show a more typical late-replication bias.

Approximately 50% of individuals possess potentially damaging exonic sSNVs in >2% of cortical cells. Every study of somatic mutation is inherently limited by tissue sampling, as somatic mutations by nature are present in some regions of particular organs but not in others. Therefore, it has traditionally been difficult to estimate body-wide or even organ-wide somatic mutation burden. We utilized our cell generation assignments to estimate global somatic mutation burden in the brain.

Assuming early embryo mutation rates as estimated above, a stable mutation rate in the fourth and the fifth cell generations, and no proliferative advantage of variants, a typical individual would amass ~86 (95% confidence interval (CI): 82–90) genome-wide sSNVs in the first to fifth cell generations after conception, each present at ≥1% VAF across the adult body (Fig. 3d). Our directly measured ~26 early embryonic sSNVs, and this estimate of 86 mutations over the first five cell divisions, are of quite similar magnitude to the 6–70 de novo germline mutations[33] per individual identified in a recent WGS family-based study, creating the possibility that function-altering mosaic mutations may contribute to variation more commonly than appreciated.

To estimate this contribution to functional variation, we focused on the ~2.2% of observed variants in our dataset that are exonic, suggesting that a typical individual would have on average ~1.9 (95% CI: 0–5) exonic sSNVs (Fig. 3e) occurring in the first five cell divisions of development. Large studies have found that ~45% of new exonic variants that have not yet undergone evolutionary selection are likely to be damaging[34,35], and among our set of 35 PCR-free exonic mutations, 37% are predicted to be most damaging with scores of either NsynD4 or LOF (Supplementary Table 6). Therefore, assuming ~37% of exonic sSNVs are potentially damaging, then ~50% of individuals would carry ≥1 damaging exonic sSNV at ≥1% VAF (≥2% of cells) (Fig. 3f). Importantly, even low-VAF mutations have

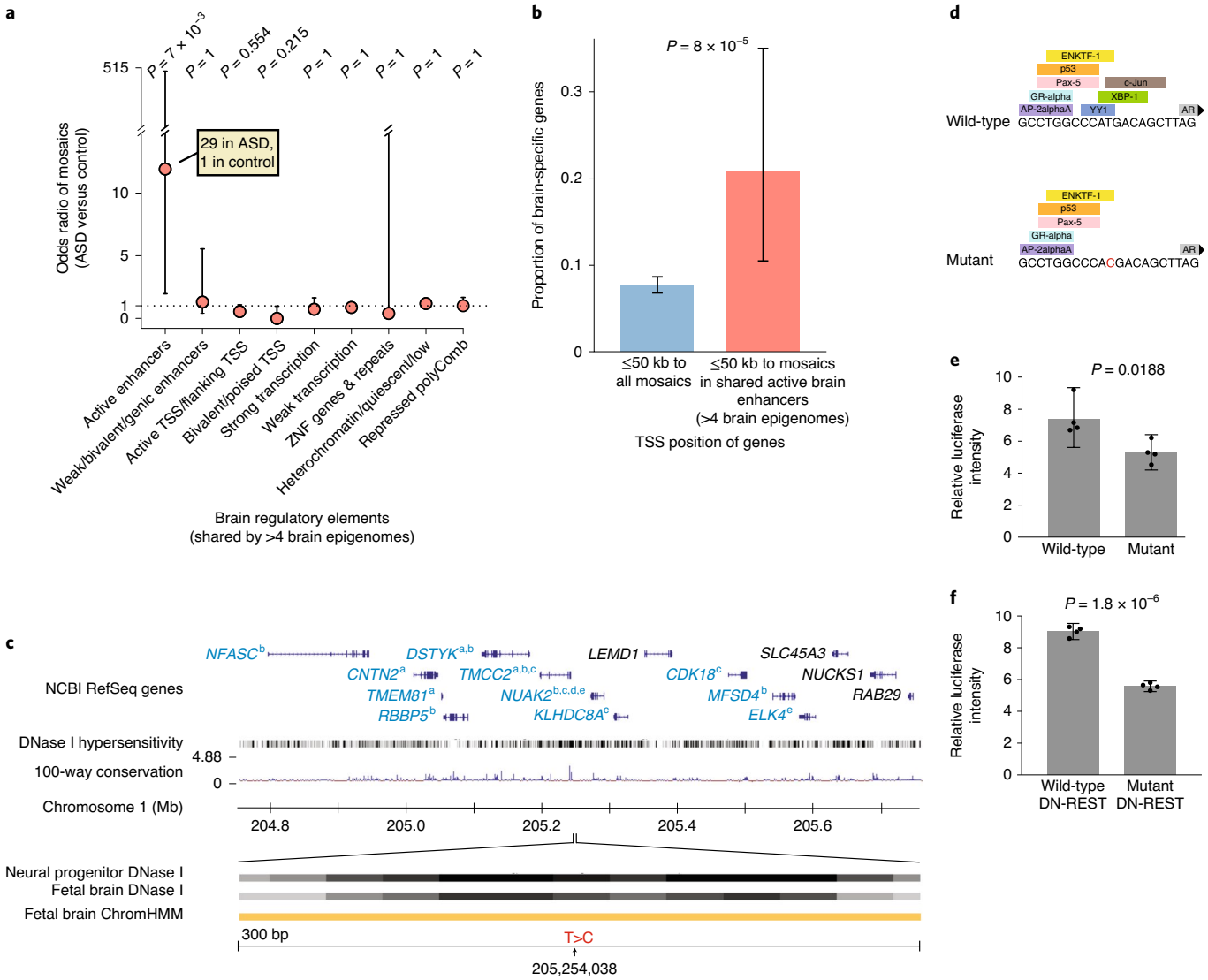

Fig. 5 | ASD brains contain somatic mutations affecting brain-active enhancers. a, Although there is no difference between ASD cases and controls in the burden of mosaic mutations present in all brain-active enhancers from the Roadmap project, ASD cases are enriched for mutations occurring in regions that act as enhancers in the majority of brain epigenomes available for analysis. Odds ratios and error bars (95% CI) were calculated by Fisher's exact test; *P* value was further corrected with Bonferroni correction. **b**, Active brain enhancer regions harboring mutations in our dataset are nearby to TSSs of genes that are enriched for brain-specific expression, compared with genes nearby to all mutations in our dataset. Error bars indicate 95% CI calculated with a binomial test. Panels **a** and **b** include sSNVs from PCR-free samples (false positive sites by deep resequencing were excluded) and validated mosaic indels, for a total of 1,689 mutations. **c**, Example of an sSNV in ASD brain AN06365 located in a brain-active enhancer. Genes in blue font have functional evidence linking their expression to enhancer activity (a, genotype tissue expression[33]; b, predicted enhancer targets[45]; c, Hi-C sequencing data[46]; d, chromatin interaction analysis by paired-end tag sequencing[47]; e, ENCODE data). The orange ChromHMM track represents active enhancer designation. **d**, The mutation is predicted to affect transcription factor binding. **e**, A mutant construct transfected into N2A cells results in reduced enhancer activity by 29% compared with wild-type construct ($n = 4$ independent plates cultured in parallel, two-tailed *t*-test, $t = 3.191$, d.f. = 6). **f**, In N2A cells pretreated with DN-REST to assume a neuronal-like state, the mutant construct reduces enhancer activity by 38% ($n = 4$ independent plates cultured in parallel, two-tailed *t*-test, $t = 18.07$, d.f. = 6). Error bars in panels **e** and **f** represent 95% CIs. Mb, megabases.

potential to cause disease, as damaging mutations with allele fractions as low as 1% and present in only a small area of the brain have been frequently reported to cause epilepsy via focal cortical dysplasia[3,4]. When we include noncoding sSNVs (as we will describe), the number of potentially damaging sSNVs increases. This analysis suggests that damaging somatic mutations are likely more common than previously appreciated, even in healthy individuals.

As an additional test, we analyzed mosaics in a second cortical region (occipital lobe, 250–300× WGS) and in nonbrain tissues (Supplementary Table 9; heart or liver, ~40–65× WGS) for two individuals in which these additional tissues were available. We found that 86% of discovered variants, usually those variants with higher VAFs arising earlier during development, were shared in at least two regions, whereas mutations with lower VAFs were often regionally restricted (Fig. 3g,h). Among the 94 sSNVs we identified in the two subjects, one is an exonic missense mutation predicted to be damaging by standard prediction tools (Methods and Supplementary Table 9), consistent with our estimate that about half of individuals will carry at least one damaging exonic somatic mutation present in a measurable fraction of cells.

Mutation types evolve in the early embryo and with replication timing. We next assessed the specific base changes in our sSNV dataset. Consistent with other reports of clonal mosaic mutation in humans[13,27,36], most sSNVs were C>T transitions (48%), with approximately half of those occurring in the context of hypermutable CpG dinucleotides. Interestingly, substitution types evolved across the first four cell divisions, with C>T transitions increasing and T>A transitions decreasing ($P=7\times10^{-6}$ and $P=2\times10^{-4}$, respectively; two-tailed Fisher's exact test) in subsequent cell generations (Fig. 4a and Supplementary Fig. 13). Although sSNVs are generally thought to be more prevalent in late-replicating regions of the genome[37,38], we observed that CpG C>T mutations were more prevalent in the earliest-replicating genomic regions, whereas non-C>T base substitutions increased in later-replicating regions, implying different DNA methylation dynamics during early embryo development[39] (Fig. 4b–d).

Autism brains contain mosaic variants affecting critical brain-active enhancers. Although statistical analysis of WGS studies is challenged by thousands of simultaneous hypotheses that can be tested, the higher rate of mutation in open chromatin that we described above suggested a specific comparison of somatic mutations with a previous study that showed a role of de novo germline mutations in neural enhancer sequences in neurodevelopmental disorders[40]. We did not observe enrichment of overall sSNVs and validated mosaic indels in brain-active enhancers in ASD cases compared with controls; however, we did observe significant enrichment when assessing only sequences bearing active enhancer marks in a majority of brain epigenomes available for analysis (from Roadmap Epigenomics[40]), reflecting those regions that are most likely to represent critical enhancers shared across individuals. When restricted to candidates that are recurrent in more than 50% (≥5 of 8) of available brain epigenomes, the odds ratio of having a mosaic mutation in an enhancer in ASD compared with that in control was 11.9 (95% CI: 1.97–487, $P=8\times10^{-4}$, two-tailed Fisher's exact test; Fig. 5a, Supplementary Figs. 14 and 15 and Supplementary Table 10). With the Bonferroni correction for testing multiple types of regulatory regions, the P value is highly significant at $P=7\times10^{-3}$. Similar enrichment was not observed in any regulatory elements active in other tissues (Supplementary Fig. 15). While this observation requires confirmation in larger datasets, our data provide a preliminary suggestion that enhancer mutations are not only especially frequent during mitotic cell divisions, but also might contribute to ASD risk in some cases.

Furthermore, genes with transcription start sites (TSSs) within 50 kilobases (kb) of these shared brain-active enhancers (≥5 brain epigenomes with active enhancer status in the Roadmap data) with mosaic mutations were enriched for brain-specific expression, implying direct functional relevance of enhancer mutations ($P=8\times10^{-5}$, two-tailed Fisher's exact test; Fig. 5b and Supplementary Table 11). While noncoding mutations in regulatory regions have been linked to ASD in other studies of peripheral blood DNA[33,40–44], our analysis, albeit limited by small sample size, shows that such mutations occur in a mosaic state in brain DNA, suggesting that perturbing the expression of brain-essential genes via brain mosaic mutation in regulatory regions could potentially contribute to ASD risk.

To investigate the functional significance of putative enhancer mutations, we engineered mutant constructs and assessed their impact on transcriptional activity via a standard luciferase activity assay in cultured neural-crest-derived cells for several variants. Case AN06365 has a validated sSNV (present in 5.3% of cells) in an enhancer on chromosome 1 near a number of important genes, some of which have functional data supporting their regulation by the enhancer[45–48] (Fig. 5c). The single-base substitution identified in this individual is predicted to impact transcription factor binding (Fig. 5d). According to our assay, this mutation has a substantial impact on transcriptional activity, resulting in reduced activity in both a neuroblastoma cell line and cells treated with dominant-negative REST (DN-REST) to produce a more differentiated neuronal state (Fig. 5e,f). Similar results were observed for a second mutation (Supplementary Fig. 16).

Discussion

In this study we investigated somatic mutations in brain DNA genome-wide using ultra-deep sequencing in a large sample of 74 brains, including 59 brains from patients with ASD. While other studies have performed targeted sequencing of ASD brain DNA[15], or whole-exome sequencing of brain DNA from other conditions[10], this represents the largest cohort of ASD brain samples to undergo deep whole-genome analysis. Our study of high-coverage WGS data has revealed a rich landscape of mutational mosaicism within both neurotypical and autistic brains, outlined rates and types of somatic mutation within the brain, and revealed insights into somatic mutation accumulation in the early embryo. We find that the first five cell divisions produce early somatic mutations in numbers comparable to de novo germline mutations, although mosaic mutations will have overall more modest effect sizes given their presence in some but not all brain cells.

Individuals vary substantially in their brain somatic mutation burden, with some brains having only a handful of detectable somatic mutations in a given region and others—even neurotypical individuals—harboring up to several dozen sSNVs present in ≥4–6% of cells, with an elevated mutation rate of sSNVs in exonic and open chromatin regions. Although somatic retrotransposon mobilization has been suggested as a major source of neuronal diversity, sSNVs formed just from the first few cell divisions number ~100 per genome (versus <1 transposon insertion per average genome[49]), in addition to several hundred more sSNVs per genome that arise later in gestation[25]. Given their relative abundance, sSNVs may contribute to interindividual genetic neural variability more than mobile elements.

Based on our estimated mutation rate per cell division, roughly half of individuals would carry one or more potentially damaging exonic mutations in a substantial fraction of cells (VAF ≥ 1%). At lower VAFs, the number of potentially function-altering somatic variants is expected to be substantially higher. Of note, predicting true deleteriousness is known to be very difficult, such that many mutations predicted as damaging would have functional impact only when homozygous and are likely to be well tolerated especially in a somatic state[50]. Still other mutations may be incompatible with life when homozygous or even when germline heterozygous, meaning that their effects could only be observed in a somatic state. Therefore, although many predicted-damaging mosaic mutations may have subtle effects depending on their distributions, they nonetheless have potential to cause or contribute to a wide number of disease states in many individuals.

Our data also suggest a potential role for mosaic mutations occurring in noncoding regulatory regions in ASD etiology, although the low availability of postmortem brains limits our sensitivity. For example, our sample size was not large enough to identify excess exonic mosaics in autism cases, although this has been documented in much larger exome studies conducted on peripheral DNA from thousands of ASD cases[11–14]. Mosaic mutations in brain-specific enhancer regions are intriguing, however, since they represent a mechanism for disrupting gene expression in brain-limited or region-specific ways, in both normal and diseased brains, without disrupting expression in other tissues. Hence, mosaic noncoding mutations represent an attractive candidate mechanism to be involved more broadly in ASD and other neuropsychiatric diseases as well.

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/s41593-020-00765-6.

Received: 10 January 2020; Accepted: 21 November 2020;
Published online: 11 January 2021

References

1. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
2. D'Gama, A. M. et al. Mammalian target of rapamycin pathway mutations cause hemimegalencephaly and focal cortical dysplasia. *Ann. Neurol.* **77**, 720–725 (2015).
3. Lim, J. S. et al. Somatic mutations in *TSC1* and *TSC2* cause focal cortical dysplasia. *Am. J. Hum. Genet.* **100**, 454–472 (2017).
4. Nakashima, M. et al. Somatic mutations in the *MTOR* gene cause focal cortical dysplasia type IIb. *Ann. Neurol.* **78**, 375–386 (2015).
5. Erickson, R. P. Recent advances in the study of somatic mosaicism and diseases other than cancer. *Curr. Opin. Genet. Dev.* **26**, 73–78 (2014).
6. Insel, T. R. Brain somatic mutations: the dark matter of psychiatric genetics? *Mol. Psychiatry* **19**, 156–158 (2014).
7. McConnell, M. J. et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: the brain somatic mosaicism network. *Science* https://doi.org/10.1126/science.aal1641 (2017).
8. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
9. Keogh, M. J. et al. High prevalence of focal and multi-focal somatic genetic variants in the human brain. *Nat. Commun.* **9**, 4257 (2018).
10. Wei, W. et al. Frequency and signature of somatic variants in 1461 human brain exomes. *Genet. Med.* **21**, 904–912 (2019).
11. Dou, Y. et al. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Hum. Mutat.* **38**, 1002–1013 (2017).
12. Freed, D. & Pevsner, J. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.* **12**, e1006245 (2016).
13. Lim, E. T. et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.* **20**, 1217–1224 (2017).
14. Krupp, D. R. et al. Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am. J. Hum. Genet.* **101**, 369–390 (2017).
15. D'Gama, A. M. et al. Targeted DNA sequencing from autism spectrum disorder brains implicates multiple genetic mechanisms. *Neuron* **88**, 910–917 (2015).
16. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched controls. *Nat. Biotechnol.* **38**, 314–319 (2020).
17. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
18. Doan, R. N. et al. Recessive gene disruptions in autism spectrum disorder. *Nat. Genet.* **51**, 1092–1098 (2019).
19. Damaj, L. et al. *CACNA1A* haploinsufficiency causes cognitive impairment, autism and epileptic encephalopathy with mild cerebellar symptoms. *Eur. J. Hum. Genet.* **23**, 1505–1512 (2015).
20. Epi4K Consortium De novo mutations in *SLC1A2* and *CACNA1A* are important causes of epileptic encephalopathies. *Am. J. Hum. Genet.* **99**, 287–298 (2016).
21. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e523 (2020).
22. Mercer, T. R. et al. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat. Genet.* **45**, 852–859 (2013).
23. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
24. Polak, P. et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
25. Ye, A. Y. et al. A model for postzygotic mosaicisms quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. *Genome Res.* **28**, 943–951 (2018).
26. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
27. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
28. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
29. Wong, C. C. et al. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat. Biotechnol.* **28**, 1115–1121 (2010).
30. Kiessling, A. A. et al. Genome-wide microarray evidence that 8-cell human blastomeres over-express cell cycle drivers and under-express checkpoints. *J. Assist. Reprod. Genet.* **27**, 265–276 (2010).
31. Gonzalez-Marin, C., Gosalvez, J. & Roy, R. Types, causes, detection and repair of DNA fragmentation in animal and human sperm cells. *Int. J. Mol. Sci.* **13**, 14026–14052 (2012).
32. Russell, L. B. & Russell, W. L. Spontaneous mutations recovered as mosaics in the mouse specific-locus test. *Proc. Natl Acad. Sci. USA* **93**, 13072–13077 (1996).
33. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
34. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
35. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
36. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
37. Chen, C. L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* **20**, 447–457 (2010).
38. 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).
39. Seisenberger, S. et al. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell* **48**, 849–862 (2012).
40. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
41. Williams, S. M. et al. An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder. *Mol. Psychiatry* **24**, 1707–1719 (2019).
42. Zhou, J. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
43. An, J. Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* https://doi.org/10.1126/science.aat6576 (2018).
44. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e712 (2017).
45. Consortium, G. T. Human genomics. The Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
46. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactome in human cells. *Proc. Natl Acad. Sci. USA* **111**, E2191–E2199 (2014).
47. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
48. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
49. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
50. Miosge, L. A. et al. Comparison of predicted and actual consequences of missense mutations. *Proc. Natl Acad. Sci. USA* **112**, E5189–E5198 (2015).

Brain Somatic Mosaicism Network

Christopher A. Walsh[1], Javier Ganz[1], Mollie B. Woodworth[1], Pengpeng Li[1], Rachel E. Rodin[1], Robert S. Hill[1], Sara Bizzotto[1], Zinan Zhou[1], Eunjung A. Lee[7], Alison R. Barton[7], Alissa M. D'Gama[1],

Alon Galor[7], Craig L. Bohrson[7], Daniel Kwon[7], Doga C. Gulhan[7], Elaine T. Lim[1], Isidro Ciriano Cortes[7], Lovelace J. Luquette[7], Maxwell A. Sherman[7], Michael E. Coulter[1], Michael A. Lodato[1], Peter J. Park[7], Rebeca B. Monroy[7], Sonia N. Kim[1], Yanmei Dou[7], Andrew Chess[8], Attila Gulyás-Kovács[8], Chaggai Rosenbluh[8], Schahram Akbarian[8], Ben Langmead[9], Jeremy Thorpe[9], Jonathan Pevsner[9], Soonweng Cho[9], Andrew E. Jaffe[10], Apua Paquola[10], Daniel R. Weinberger[10], Jennifer A. Erwin[10], Jooheon H. Shin[10], Richard E. Straub[10], Rujuta Narurkar[10], Alexej S. Abyzov[11], Taejeong Bae[11], Anjene Addington[12], David Panchision[12], Doug Meinecke[12], Geetha Senthil[12], Lora Bingaman[12], Tara Dutka[12], Thomas Lehner[12], Laura Saucedo-Cuevas[13], Tara Conniff[13], Kenneth Daily[14], Mette Peters[14], Fred H. Gage[15], Meiyan Wang[15], Patrick J. Reed[15], Sara B. Linker[15], Alex E. Urban[16], Bo Zhou[16], Xiaowei Zhu[16], Aitor Serres[17], David Juan[17], Inna Povolotskaya[17], Irene Lobón[17], Manuel Solis-Moruno[17], Raquel García-Pérez[17], Tomas Marquès-Bonet[17], Gary W. Mathern[18], Eric Courchesne[19], Jing Gu[19], Joseph G. Gleeson[19], Laurel L. Ball[19], Renee D. George[19], Tiziano Pramparo[19], Diane A. Flasch[20], Trenton J. Frisbie[20], Jeffrey M. Kidd[20], John B. Moldovan[20], John V. Moran[20], Kenneth Y. Kwan[20], Ryan E. Mills[20], Sarah B. Emery[20], Weichen Zhou[20], Yifan Wang[20], Aakrosh Ratan[21], Michael J. McConnell[21], Flora M. Vaccarino[22], Gianfilippo Coppola[22], Jessica B. Lennington[22], Liana Fasching[22], Nenad Sestan[22] and Sirisha Pochareddy[22]

[7]Harvard University, Boston, MA, USA. [8]Icahn School of Medicine at Mt Sinai, New York, NY, USA. [9]Kennedy Krieger Institute, Baltimore, MD, USA. [10]Lieber Institute for Brain Development, Baltimore, MD, USA. [11]Mayo Clinic, Rochester, MN, USA. [12]NIMH, Bethesda, MD, USA. [13]Rockefeller University, New York, NY, USA. [14]Sage Bionetworks, Seattle, WA, USA. [15]Salk Institute for Biological Studies, La Jolla, CA, USA. [16]Stanford University, Stanford, CA, USA. [17]Universitat Pompeu Fabra, Barcelona, Spain. [18]University of California, Los Angeles, CA, USA. [19]University of California, San Diego, CA, USA. [20]University of Michigan, Ann Arbor, MI, USA. [21]University of Virginia, Charlottesville, VA, USA. [22]Yale University, New Haven, CT, USA.

Methods

Human tissue and DNA samples. For detailed information on experimental design, please refer to the Nature Research Reporting Summary. Frozen postmortem human brain specimens from 61 ASD cases and 15 neurotypical controls were obtained from the Lieber Institute for Brain Development and the University of Maryland through the NIH NeuroBioBank, as well as from Autism BrainNet. All specimens were deidentified and all research was approved by the institutional review board of Boston Children's Hospital. No statistical methods were used to predetermine sample sizes but we attempted to obtain samples from all donated postmortem ASD brains with frozen prefrontal cortex available for research, as well as a reasonable number of control brains. Data collection and analysis were not performed blind to the conditions of the experiments. Data collection was not randomized.

DNA was extracted from dorsolateral prefrontal cortex where available (or generic cortex in a minority of cases) using lysis buffer from the QIAamp DNA Mini kit (Qiagen) followed by phenol chloroform extraction and isopropanol cleanup. Samples UMB4334, UMB4899, UMB4999, UMB5027, UMB5115, UMB5176, UMB5297, UMB5302, UMB1638, UMB4671 and UMB797 were processed at New York Genome Center using TruSeq Nano DNA library preparation (Illumina) followed by Illumina HiSeq X Ten sequencing to a minimum of 200× depth. All remaining samples were processed at Macrogen using TruSeq DNA PCR-Free library preparation (Illumina) followed by minimum of 30× sequencing of seven separate libraries on the Illumina HiSeq X Ten, for a total minimum coverage of 210× per sample. We achieved an average of 251× depth across all samples, using 150-base-pair (bp) paired-end reads. Two samples, UMB5771 and UMB5939, had parental saliva-derived DNA available, and DNA from both parents for these two cases was obtained and sequenced at Macrogen to ~50× depth. Parental DNA was not available for any other samples. Additionally, DNA was extracted from Brodmann Area 17 (occipital lobe) for cases UMB4638 and UMB4643 and sequenced at Macrogen to a minimum of 210× depth following PCR-free library preparation. Bulk heart and liver sequencing data, as well as single-cell sequencing data from three individuals (UMB1465, UMB4643 and UMB4638), were previously published by our group and used again in this study[8,17].

Mutation calling and filtration. All paired-end FASTQ files were aligned using BWA-MEM v.0.7.8 to the GRCh37/hg19 human reference genome including the hs37d5 decoy sequence from the Broad Institute, following GATK best practices[51,52]. We used MuTect2-PoN[53] (GATK v.3.5) to generate a set of PoNs (panel-of-normals), by using 73 individuals other than the sample being analyzed (including both cases and controls), to remove sequencing artifacts and germline variants. For somatic mutation calling, rare variants were further selected by filtering out any variant with a maximum population minor allele frequency (MAF) $>1\times10^{-5}$ in the Genome Aggregation Database (gnomAD)[54]. Variants within segmental duplication regions or nondiploid regions[16] were also removed. Low-quality calls tagged 't_lod_fstar', 'str_contraction' and 'triallelic_site' were removed. A minimum VAF of 0.03 was required unless a variant was phasable by MuTect2, which allowed for rescue of variants down to VAF of 0.02; however, a threshold of 0.03 was maintained for PCR-based samples. A minimum alternative read depth of three reads was required. Only private events among the population were analyzed. An upper VAF threshold of 0.40 was set and heterozygous germline variants were removed. For mosaic indels specifically, variants within RepeatMasker regions (http://www.repeatmasker.org/) and simple repeats regions[55] were further excluded.

We then used MosaicForecast[16] to perform read-backed phasing and identify high-confidence mosaics from the candidate call set. Briefly, features likely to be correlated with mosaic detection specificity were selected: mapping quality, base quality, clustering of mutations, read depth, number of mismatches per read, read1/read2 bias, strand bias, base position, read position, trinucleotide context, sequencing cycle, library preparation method and genotype likelihood. Based on these features, a random forest model was trained using phased variants. Further training was conducted using parental WGS data from two cases, UMB5771 and UMB5939, as well as single-cell WGS data from three control brains, UMB1465, UMB4643 and UMB4638 (refs. [8,17]), for which we constructed lineage trees with the sSNVs we identified and assigned variants to different clades, and germline variants were identified as those presenting in multiple conflicting clades[16]. Predicted mosaics were further filtered by removing genomic regions enriched for low-VAF variants and by removing variants with unusually high sequencing depth that also occurred in regions marked as copy number variants by Meerkat[56]. Following all training and filtration, we identified 2,166 putative mosaic sSNVs (Supplementary Table 2). Two ASD samples, ABN_B6S3 and UMB5308, were eliminated from the study at this stage due to very high noise suggestive of sample contamination, leaving 59 ASD cases with high-quality sequencing data.

Prediction of pathogenicity scores. Pathogenicity prediction scores were calculated for functional mosaic and germline variants using a modified version of a previously described pipeline[18]. The pipeline uses 12 different prediction tools (SIFT, LRT, MutationTaster, MutationAssesor, FATHMM, Provean, MetaSVM, MetaLR, M-CAP, MutPred, Eigen and CADD) and classifies variants as follows: Nsyn, missense variant with ≥5 benign predictions or 0 damaging predictions; NsynD1, damaging missense variant with ≥1 damaging prediction and <5 benign predictions; NsynD2, damaging missense variant with ≥4 damaging predictions and <5 benign predictions; NsynD4, damaging missense variant with ≥5 damaging predictions and <5 benign predictions and genomic evolutionary rate profiling score (GERP) > 2 (CADD > 15, DANN > 0.9, EIGAN > 0.9, REVEL > 0.9); LOF-1, stopgain/frameshift; LOF-2, canonical splicing (intronic ± 1–2 bases); LOF-3, exonic splicing sites ±2 bp or intronic splicing region (±3–15 bp) plus splicing impact prediction; LOF-4, other sites with large splicing prediction; and LOF-5, stop-loss, likely benign mutations in splicing regions, extended splicing, not predicted to cause change and GERP < 2. Gene constraint was calculated with pLI scores[57] and with missense and synonymous Z scores[58]. Loss-of-function was also assessed with LOFTEE analysis[59]. Genes were also screened through the Online Mendelian Inheritance in Man (OMIM) database of genes with relevance to any human disease (http://www.omim.org/).

Germline mutation analysis. Germline mutations were classified using the following criteria: private variants in exome or annotated splice sites, not in low-complexity regions identified by RepeatMasker (http://www.repeatmasker.org), alternative allele depth ≥4 reads, VAF ≥ 0.03, gnomAD v.2.1.1 (ref. [54]) population MAF < 0.001% or allele count ≤5, and predicted to be heterozygous by MosaicForecast. Following the results of a recent large analysis of germline coding variation in ASD and control exomes[21], we filtered germline protein-truncating variants on the basis of pLI score (pLI ≥ 0.995 (ref. [58])), and germline missense variants on the basis of missense badness/Polyphen2/constraint score (MPC ≥ 2 (ref. [60])). Both of these filters enrich for variants that are depleted in population reference panels, and variants that meet these criteria were shown to be in excess in ASD versus control exomes[21]. To account for the relatively weaker degree of risk conferred by high-MPC missense variants in case–control designs (risk ratio = 1.2 (ref. [21])), we further filtered for high-MPC variants in genes that met genome-wide significance in the Autism Sequencing Consortium analysis. In summary, our category of likely pathogenic germline variants includes protein-truncating variants in high-pLI genes, as well as high-MPC missense variants in the 102 genes identified as important ASD genes by Satterstrom et al.[21]. We tested for a burden of such variants in ASD versus control brains with a Poisson exact test, following the Autism Sequencing Consortium analysis.

Amplicon resequencing validation. Targeted validation was attempted on 208 sSNVs and all called indels. Additional validation was conducted on called exonic sSNVs that were ultimately excluded from the dataset due to low VAF in PCR-based samples or presence in gnomAD database (Supplementary Table 6). Validation candidates were selected based on potential functional significance, ability to design PCR primers, and representative diversity of VAFs and genomic loci. Multiple sets of PCR primers were designed for each variant and synthesized with Ion Torrent adapters P and A, with barcodes added for unique identification. PCR amplification was performed using Phusion HotStart II DNA Polymerase (Thermo) as described by the manufacturer, with 20–25 cycles of amplification. Reactions were pooled and purified with AMPure XP technology (Agencourt), then sequenced on the Ion Torrent Personal Genome Machine using the Ion 530 chip with 400-bp reads, reaching an average coverage of 92,000 reads per variant, amongst sSNV reactions that yielded mappable reads.

Following demultiplexing and trimming, reads were mapped using BWA-MEM and locally realigned using GATK. High-quality reference and mutant reads were then counted using mpileup and variants with successful PCR reactions resulting in usable reads were then classified as validated true mosaics or homozygous reference with variant not present. Any ambiguous variants, including variants in which there was discordance between sequencing from different PCR primers, were conservatively assigned a designation of homozygous reference. Validation success rates were calculated as the number of true mosaics divided by the sum of true mosaics and homozygous reference, excluding variants from brains UMB1465, UMB4643 and UMB4638 as validation in these brains was conducted on an alternative DNA source as none of the originally sequenced DNA remained. Weighted averaging across PCR and PCR-free variant validation was used to determine a comprehensive validation rate of 90%. Five variants from UMB5771 and UMB5939 were also resequenced in parental DNA, which confirmed a mosaic state in the offspring and homozygous reference in parents.

Epigenetic covariates of mosaic mutations. Candidate mosaic mutations were annotated with ANNOVAR[61] (v.2017-07-17) to calculate the observed density of putative mosaics in different regions. Rare single-nucleotide polymorphisms (SNPs) (MAF < 0.01) from 15,708 whole genomes of unrelated individuals in gnomAD[58] were annotated with ANNOVAR and used to calculate the expected density of mutations in different regions. DNase I hypersensitive regions for different tissues were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/DNase/, and the DNAse I-accessible regulatory regions (FDR 0.01) were used to calculate the in- and out-of-region density of putative mosaics. We merged DHS regions in different tissues profiled by the Roadmap Epigenomics Project to obtain a single set of DHS regions.

Chromatin states in different tissues and cell lines predicted by Hidden Markov Model v.1.10 using 18 states (6 marks, 98 epigenomes) were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/ (ref. [62]). Mutations in patients with ASD versus control samples in different regulatory regions were compared using a two-tailed Fisher's exact test. A Bonferroni correction for multiple hypothesis testing was implemented for nine comparisons as specified below. Roadmap epigenomes were separated for analysis based on their tissue of origin. States were classified as follows: 9_EnhA1 and 10_EnhA2, active enhancers; 7_EnhG1, 8_EnhG2, 11_EnhWk and 15_EnhBiv, weak/bivalent/genic enhancers; 1_TssA, 2_TssFlnk, 3_TssFlnkU and 4_TssFlnkD, active TSS/flanking TSS; 14_TssBiv, bivalent/poised TSS; 5_Tx, strong transcription; 6_TxWk, weak transcription; 12_ ZNF/Rpts, ZNF genes and repeats; 13_Het and 18_Quies, heterochromatin/quiescent/low; and 16_TssBiv and 17_ReprPCWk, repressed polycomb.

Simulation of mosaic mutations and calculation of sensitivity. The 300× WGS data for NA12878 (Genome in a Bottle, downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab) were downsampled to 250× using SAMtools[63]. High-confidence SNP calls for individual NA12878 were downloaded from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.18/ (ref. [64]). Simulated mosaic mutations with different VAFs were generated in the 250× BAM file by converting bases supporting the alternative alleles of high-confidence heterozygous SNPs to reference bases at several binomial sampling probabilities. Simulated sites with expected VAFs of 0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.1, 0.15, 0.2, 0.25, 0.3, 0.32, 0.34, 0.36, 0.38, 0.4, 0.45 and 0.5 were generated and used to calculate sensitivities. The 95% CIs for sensitivity at each VAF were calculated using a binomial test.

Symmetric versus asymmetric cell contribution analysis. Ion Torrent amplicon resequencing for 96 germline heterozygous mutations revealed that VAFs were over-dispersed compared with a binomial distribution (Supplementary Fig. 3), likely due to noise induced by PCR amplification as part of resequencing. We fit the VAF distribution with a beta-binomial model to capture the over-dispersion (aggregation parameter $\theta = 278$, correlation parameter $\rho = \frac{1}{1+\theta} = 0.0036$). An R package, VGAM[65] (v.1.1-1), was used to fit the beta-binomial model. It has previously been reported that the two ancestor cells of the blood lineage give rise to offspring asymmetrically at approximately a 2:1 ratio[26]. Here we used Ion Torrent-validated mosaics from diploid chromosomes with a similar model to measure potential asymmetric cell contributions to the brain during early embryonic development. Briefly, we let α_1 and $1-\alpha_1$ be the fraction of brain cells deriving from each of the two cells created by the first division of the brain ancestor cell. A contribution parameter value of $\alpha_1 = 0.5$ means the first two cells contributed equally to the brain, while a non-0.5 value means the cell contribution is asymmetrical. Given a specific α_1, it is possible to calculate the expected VAF for mutations acquired at different branches of the early phylogeny. Assuming the mutation rate per cell generation is constant (that is, the two cell divisions from the second cell generation have the same mutation rate), we compute the likelihood of a mosaic arising on a specific branch by multiplying the estimated sensitivity for detecting mosaics at the expected branch VAF with the over-dispersion beta-binomial likelihood of the mosaic VAF measured by the deep Ion Torrent sequencing. The log likelihoods for all sites were then summed over all branches to estimate the log likelihood of a specific α_1. We fit α_1 by maximizing the log likelihood over $\alpha_1 \in [0.5,1]$ using a grid search with step size = 0.0002. A likelihood ratio test was used to compare the asymmetric model with the symmetric model ($\alpha_1 = 0.5$), which favored the model with unequal cell contribution during the first cell generation ($P = 3\times10^{-4}$). A 95% CI for α_1 (0.555, 0.597; Supplementary Fig. 11) was constructed using the likelihood ratio (all values of α_1 for which the likelihood drops off by no more than 1.92 units).

To examine the potential VAF dispersion problem in our 250× WGS data, we randomly extracted 25,000 phasable germline sites with 0.4–0.6 VAF (by MuTect2) and plotted the VAF distribution profile. No VAF over-dispersion was found compared with binomial sampling (Supplementary Fig. 3).

Mutation rate estimation and assignment of mutations to cell generations. To estimate per-generation mutation rates, we used an expectation-maximization algorithm similar to that described by Ju et al.[26]. Briefly, the mutation rate ν_g for all cell generations was considered to be identical at the beginning, and the probability of a mutation j belonging to cell generation g (expectation step) is

$$p_{g,j} = \frac{\sum_{b\in b_g} P\left(\mathrm{VAF}_b, \mathrm{dp}_j, \mathrm{alt}_j\right) s_b \nu_b}{\sum_{b=1}^{B} P\left(\mathrm{VAF}_b, \mathrm{dp}_j, \mathrm{alt}_j\right) s_b \nu_b},$$

where $P(\mathrm{VAF}_b, \mathrm{dp}_j, \mathrm{alt}_j)$ is the binomial probability of observing alt_j successes (alternative allele-supporting reads) in dp_j trials (total reads) with probability of success VAF_b. B denotes the total number of branches for the first to fifth cell generations; ν_b is the mutation rate for branch b; b_g is the set of all branches belonging to generation g; and s_b denotes the sensitivity for detecting mosaics on branch b. We assumed the same mutation rate for all branches in a specific generation, and symmetric contributions of cells to the embryo. The mutation rate for cell generation g was then updated as the sum of $p_{g,j}$ across all mosaics (maximization step):

$$N_g = \sum_{j=1}^{N} p_{g,j}$$

The two steps were iterated until convergence.

To obtain upper and lower bounds for mutation rate per cell generation, we ran several different bootstrap simulations over mosaics from the 63 PCR-free samples. Bootstrap resamplings (sampling 63 brains each time) were performed 1,000 times to estimate the distribution of observed mutations per cell division (Supplementary Fig. 8), and for each cell division the total number of mosaics was obtained by dividing the observed number of mosaics by the VAF-specific detection sensitivities. A 95% CI was computed by calculating 2.5–97.5% percentiles.

The mutation rate for the fourth cell generation and the mutation rate for the fifth cell generation were added up to give an estimate of the mutation rate for 4+ cell generations, and the mutation rates per cell division for the first, second, third and 4+ cell generations were estimated to be 3.37, 2.51, 2.28 and 2.85, respectively. Using our WGS data, the probability of each mutation belonging to each cell generation was obtained, and all 1,641 putative mutations from PCR-free samples and Ion Torrent-validated PCR-based samples were then assigned to different cell generations using maximum likelihood values. In total, 182 mutations were assigned to the first cell generation, 250 to the second cell generation, 392 to the third cell generation and 817 to the fourth cell generation and beyond (Supplementary Table 8). We also compared mutation-to-generation assignments determined by deep WGS bulk data with those calculated from single-cell data for three previously analyzed individuals[8,17] and found reasonable agreement between the two methods (Fig. 3b and Supplementary Fig. 12). Our cell generation assignments did not change appreciably when based on asymmetric versus symmetric models of cell division (98.6% in concordance; Supplementary Table 8).

Estimation of total mutations and exonic mutations per individual. There are 31 cell divisions ($2^0+2^1+2^2+2^3+2^4$) in the first five cell generations of early embryonic development. Mutations per cell division for the first to fifth cell generations were bootstrapped from the values we generated in the section above, and the total number of mutations was calculated by adding up all mutations from the first to fifth cell generations. The process was repeated 10,000 times to estimate the total number of mutations per individual. We calculated the number of exonic mutations using our data that 2.2% of called sSNVs were exonic. Reported values were obtained by simulating binomial sampling on the total number of mutations for each individual. A 95% CI was computed by calculating 2.5–97.5% percentiles.

Comparison of mosaics in brain and nonbrain tissues. Mosaic mutations were called from WGS data for two different brain regions (PFC and BA17, representing occipital lobe) in two individuals (UMB4643 and UMB4638) using the same pipeline described above. All putative mosaic mutations from each region were visually inspected by SAMtools mpileup across different tissues, including the two brain regions with ~250× depth of coverage and one nonbrain tissue (liver or heart) with ~50× depth of coverage. A mutation was considered to be absent from the tissue if there were no alternative allele-supporting reads observed in that tissue.

Comparison of mutational signatures between earlier and later mutations. We downloaded all 96-dimensional mutational signatures from PCAWG[66]. To avoid over-fitting, we extracted the two most common clock-like signatures (signature S1 and signature S5) as well as a signature highly related with sequencing artifacts (signature S18) from the PCAWG signatures, and deconstructed mutational signatures for the mosaic mutations using the R package deconstructSigs[67]. We observed a trend toward an increase in signature S1 across the first to fourth cell generations (Supplementary Fig. 13), which is believed to be caused by an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine (ref. [68]). Mutation profiles from different cell generations were compared using a two-tailed Fisher's exact test.

Comparison of DNA replication timing between mutations from different cell generations. We extracted locus-specific DNA replication timing for all putative mosaic mutations using AsymTools (http://software.broadinstitute.org/cancer/cga/AsymTools)[69]. The sSNVs were then classified into four categories according to DNA replication time quartiles.

Evaluation of gene expression level with TSSs near shared brain-active enhancers. TSSs of coding genes were extracted from GENCODE[70] v.19 annotations on GRCh37. Genes with their TSSs overlapping regions within 50 kb upstream or downstream of sSNVs from PCR-free samples were extracted. Tissue-specific expression derived for a total of 53 tissues and cell types was downloaded from the GTEx project[71]. We used the expression table from GTEx v.7 (gene median transcripts per million per tissue), and brain-specific genes were defined as: (1) genes with median expression level across brain tissues at least three times higher than the median expression level across all tissues; and (2) genes in which the highest median expression level in any tissue occurs in a brain tissue. A two-tailed Fisher's exact test was applied to compare types of genes (brain-specific

genes or others) near sSNVs (genes with TSSs adjacent to mosaics in shared brain-active enhancers versus genes with TSSs adjacent to all mosaic sites).

Luciferase assays for assessment of enhancer activity. We selected 17 sSNVs identified in brain-active enhancers and attempted cloning and site-directed mutagenesis (New England Biolabs) to recreate the mutations. Mutagenesis was successful for 11 mutations. Wild-type and mutant constructs were then cloned into luciferase vector pGL4.25 (Promega). Luciferase plasmids were transfected into N2A cells, along with an internal control plasmid (phRL-TK(Int-), Promega) and DN-REST[72] or GFP expression plasmids using Lipofectamine LTX with PLUS reagent (Thermo Fisher). Luciferase activities were measured 24 h later. Three wild-type enhancer constructs had technically successful assays with significant difference from negative control, and, among these, two showed a significant difference between wild type and mutant. All experiments were performed with $n = 4$ and averaged across replicates.

Statistical analysis. All data were graphed and analyzed using R (v.3.6.1) and Python (v.3.6.7). Group differences were calculated with two-tailed Wilcoxon rank sum test (Fig. 2f), permutation (Fig. 3c) or two-tailed Student's *t*-test (Fig. 5e,f). For the data tested with Student's *t*-test in Fig. 5e,f, data distribution was assumed to be normal but this was not formally tested. All of the error bars in our manuscript represent 95% CIs, except for boxplots. In the boxplots, the lower and upper hinges correspond to the first and third quartiles, and the middle lines correspond to the median values. Binomial proportion confidence interval was used to estimate the 95% CIs of alternative allele fractions (Fig. 2d), the observed mutation rate across different regions (Fig. 2e), different fractions of base substitutions (Fig. 4a) and proportions of brain-specific genes with their TSSs near mosaic mutations (Fig. 5b). A two-tailed Fisher's exact test was used to calculate whether the proportions of one variable are different depending on the value of other variables (Figs. 4a and 5a,b).

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

Data availability

Whole-genome sequencing data are available from the National Institute of Mental Health Data Archive (https://doi.org/10.15154/1503337).

Code availability

Custom code is available from the authors by request.

References

51. Genovese, G., Handsaker, R. E., Li, H., Kenny, E. E. & McCarroll, S. A. Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. *Am. J. Hum. Genet.* **93**, 411–421 (2013).
52. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
54. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
55. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
56. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
57. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
58. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
59. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
60. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353v1 (2017).
61. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
62. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
63. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
65. Yee, T. W. *Vector Generalized Linear and Additive Models: with an Implementation in R* (Springer-Verlag, 2015).
66. Alexandrov, L. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
67. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
68. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–4217 (2013).
69. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
70. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
71. Consortium, G. T. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
72. Chong, J. A. et al. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).

Acknowledgements

Human tissue was obtained from the NIH NeuroBioBank at the University of Maryland, the Lieber Institute for Brain Development, Oxford University Brain Bank and Autism BrainNet. Autism BrainNet is a resource of the Simons Foundation Autism Research Initiative (SFARI). Autism BrainNet also manages the Autism Tissue Program (ATP) collection previously funded by Autism Speaks. We thank the donors and their families for their invaluable contribution to the advancement of science. We also thank R.S. Hill, J. Partlow, W. Bainter and the Research Computing group at Harvard Medical School for assistance. R.E.R., A.M.D. and S.N.K. are supported by the Stuart H.Q. and Victoria Quan Fellowship in Neurobiology. R.E.R., A.M.D. and A.N. are also supported by the Harvard/MIT MD-PhD program (grant no. T32GM007753) from the National Institute of General Medical Sciences. Y.D., M.K., M.A.S., D.C.G. and P.J.P. are supported by grants from the NIMH (grant nos. U01MH106883 and P50MH106933) and the Harvard Ludwig Center. L.J.L. and C.L.B. are supported by the Bioinformatics and Integrative Genomics training grant (no. T32HG002295) from the National Human Genome Research Institute. C.A.W. is supported by the Manton Center for Orphan Disease Research, the Allen Discovery Center program through The Paul G. Allen Frontiers Group, grant no. R01NS032457 from the NINDS and grant no. U01MH106883 from the NIMH. C.A.W. is an Investigator of the Howard Hughes Medical Institute. Data were generated as part of the Brain Somatic Mosaicism Network (BSMN) Consortium, supported by grant nos.: U01MH106874, U01MH106876, U01MH106882, U01MH106883, U01MH106883, U01MH106884, U01MH106891, U01MH106891, U01MH106891, U01MH106892, U01MH106893 and U01MH108898 awarded to: N.S. (Yale University), F.M.V. (Yale University), F.H.G. (Salk Institute for Biological Studies), C.A.W. (Boston Children's Hospital), P.J.P. (Harvard University), J.P. (Kennedy Krieger Institute), A.C. (Icahn School of Medicine at Mount Sinai), J.V.M. (University of Michigan), D.R.W. (Lieber Institute for Brain Development) and J.G.G. (University of California, San Diego). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

Author contributions

R.E.R., Y.D., P.J.P. and C.A.W. conceptualized the study. R.E.R., A.M.D. and S.N.K. generated whole-genome sequencing data. Y.D. led bioinformatic analysis with assistance from M.K. for variant identification and from M.A.S., L.J.L., C.L.B. and D.C.G for technical issues. R.E.R., L.M.R. and R.N.D. performed targeted variant validation. L.M.R. and K.M.G. performed cloning and luciferase assay experiments. A.N. performed germline pathogenic variant analysis. R.E.R., Y.D., P.J.P. and C.A.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-00765-6.

Correspondence and requests for materials should be addressed to P.J.P. or C.A.W.

Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

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

nature research

Corresponding author(s): Peter J. Park, Christopher A. Walsh

Last updated by author(s): May 8, 2020

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 our Editorial Policies 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 data in this study.
Data analysis	We aligned FASTQ files using BWA-MEM version 0.7.8. We used MuTect2 with panel of normals from GATK version 3.5 to call variants. We identified somatic variants using a custom method called MosaicForecast, which is published separately. Our analysis also used RepeatMasker version 4.0.7, Meerkat version 0.189, ANNOVAR version 2017-07-17, LOFTEE version 1.0.3, CADD version 1.4, SIFT 4G, LRT (original version), MutationTaster2, MutationAssesor release 3, FATHMM version 2.3, Provean version 1.1.3, MetaSVM version 1.0, MetaLR (original version), M-CAP version 1.4, MutPred2, Eigen version 1.1, VGAM version 1.1-1, deconstructSigs version 1.8, SAMtools-1.10, and AsymTools (original version).

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 and 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

Whole-genome sequencing data is available from the National Institute of Mental Health Data Archive (DOI: 10.15154/1503337).

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	An attempt was made at procuring all donated autism brains in the world for which fresh-frozen prefrontal cortex was available. In total, 61 ASD brains were obtained. Due to the high cost of ultra-deep sequencing and importance of focusing resources on autism cases, 15 matched controls were selected and sequenced. These sample sizes are small compared to studies on peripheral autism DNA, but the sample sizes are very large in comparison to prior WGS studies of human brain.
Data exclusions	Two autism samples was excluded from analysis due to likely sequencing contamination. Extremely high rates of mutation calling and extremely low validation success rate were consistent with contamination at the level of sequencing. These criteria were not established prior to data analysis, but the inadequacy and contamination of these samples was abundantly obvious.
Replication	Deep re-sequencing of a large set of putative mosaic mutations was attempted, and confirmed a very high validation success rate. Protein damaging effects were also determined via multiple published softwares. These methods ensured replication of results.
Randomization	This is not relevant to our study. ADI-R scores were obtained whenever possible in order to verify that ASD cases did indeed have autism, and communication with brain banks confirmed that controls were neurologically normal.
Blinding	Blinding was not relevant to this study. In order to obtain samples from brain banks, it was necessary to specify and know disease status, and this information was known to all researchers throughout the study.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Neuro-2a cells from ATCC CCL-131
Authentication	Cell lines were not authenticated
Mycoplasma contamination	Regular mycoplasma testing was conducted during culture of cells used in this study and was negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.