nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-022-01180-2

Check for updates

Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements

Lovelace J. Luquette[1,16], Michael B. Miller[2,3,4,5,6,16], Zinan Zhou[2,16], Craig L. Bohrson[1], Yifan Zhao[1], Hu Jin[1], Doga Gulhan[1], Javier Ganz[2], Sara Bizzotto[2], Samantha Kirkham[2], Tino Hochepied[7,8], Claude Libert[7,8], Alon Galor[1], Junho Kim[2,9], Michael A. Lodato[10], Juan I. Garaycoechea[11], Charles Gawad[12,13], Jay West[14], Christopher A. Walsh[2,3,15,17] ✉ and Peter J. Park[1,17] ✉

Accurate somatic mutation detection from single-cell DNA sequencing is challenging due to amplification-related artifacts. To reduce this artifact burden, an improved amplification technique, primary template-directed amplification (PTA), was recently introduced. We analyzed whole-genome sequencing data from 52 PTA-amplified single neurons using SCAN2, a new genotyper we developed to leverage mutation signatures and allele balance in identifying somatic single-nucleotide variants (SNVs) and small insertions and deletions (indels) in PTA data. Our analysis confirms an increase in nonclonal somatic mutation in single neurons with age, but revises the estimated rate of this accumulation to 16 SNVs per year. We also identify artifacts in other amplification methods. Most importantly, we show that somatic indels increase by at least three per year per neuron and are enriched in functional regions of the genome such as enhancers and promoters. Our data suggest that indels in gene-regulatory elements have a considerable effect on genome integrity in human neurons.

Although somatic mutation has been studied extensively in cancer, investigation into the abundance, patterns and effects of somatic mosaicism in nonneoplastic tissues has only recently begun[1–6]. Unlike tumor tissue in which somatic mutations of interest are shared by large clones, somatic mutations in normal tissues are typically shared by relatively few cells and are hence difficult to detect. Recent studies have circumvented the technical difficulty of detecting rare somatic mutations by ultradeep sequencing of very small tissue samples[3,7], exploiting naturally occurring genetically homogeneous clones[8] or clonal expansion of cells in vitro[5,9,10].

Another strategy for detecting somatic mosaic mutations is to directly sequence DNA from a single cell. Single-cell DNA sequencing (scDNA-seq) is capable of detecting the rarest somatic mutations (that is, mutations private to a single cell) and can also provide information about cell lineage through shared somatic mutations[2,11]. This strategy is especially useful for examining postmitotic cells such as neurons. A major challenge, however, is the difficulty of amplifying the genome of a single cell accurately and evenly before sequencing. For example, multiple displacement amplification (MDA)[12], a popular amplification method for detecting point mutations, produces nonuniformity across the genome[13] and often amplifies homologous alleles of diploid cells at different rates, leading to allelic imbalance[14]. These amplification artifacts pose substantial challenges for identifying mutations from short-read sequencing data—especially mutations that are nonclonal and thus cannot be confirmed in other cells. We previously used LiRA[15], a tool based on read-level phasing, to filter artifacts in MDA samples and discovered an age-associated increase in somatic mutations in human neurons[6]; however, this approach was limited to analyzing mutations within a few hundred base-pairs of SNPs, making it adequate for estimating the overall mutational spectrum and burden in a sample, but not for other analyses. Another method, SCAN-SNV[14], could find SNVs over more of the genome by estimating local allelic imbalance, but it was optimized for MDA-amplified data.

A new single-cell amplification method known as primary template-directed amplification (PTA) reduces amplification-associated artifacts by dampening the exponential nature of isothermal MDA[16]. Indeed, our comparison below of single neurons amplified by both the MDA and the PTA protocols from the prefrontal cortices (PFCs) of the same individuals shows that PTA substantially improves on MDA. Despite PTA's improvements, the resulting data still require specialized single-cell mutation calling, as conventional bulk-oriented somatic SNV (sSNV) analysis based on the Genome Analysis Toolkit's (GATK) best practices can yield an

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [3]Manton Center for Orphan Disease, Boston Children's Hospital, Boston, MA, USA. [4]Division of Neuropathology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA. [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Center for Inflammation Research, Flanders Institute for Biotechnolpogy (VIB), Ghent, Belgium. [8]Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium. [9]Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea. [10]Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA, USA. [11]Hubrecht Institute—KNAW, University Medical Center Utrecht, Utrecht, the Netherlands. [12]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [13]Chan Zuckerberg Biohub, San Francisco, CA, USA. [14]BioSkryb, Durham, NC, USA. [15]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA. [16]These authors contributed equally: Lovelace J. Luquette, Michael B. Miller, Zinan Zhou. [17]These authors jointly supervised this work: Christopher A. Walsh, Peter J. Park. ✉e-mail: Christopher.Walsh@childrens.harvard.edu; peter_park@hms.harvard.edu

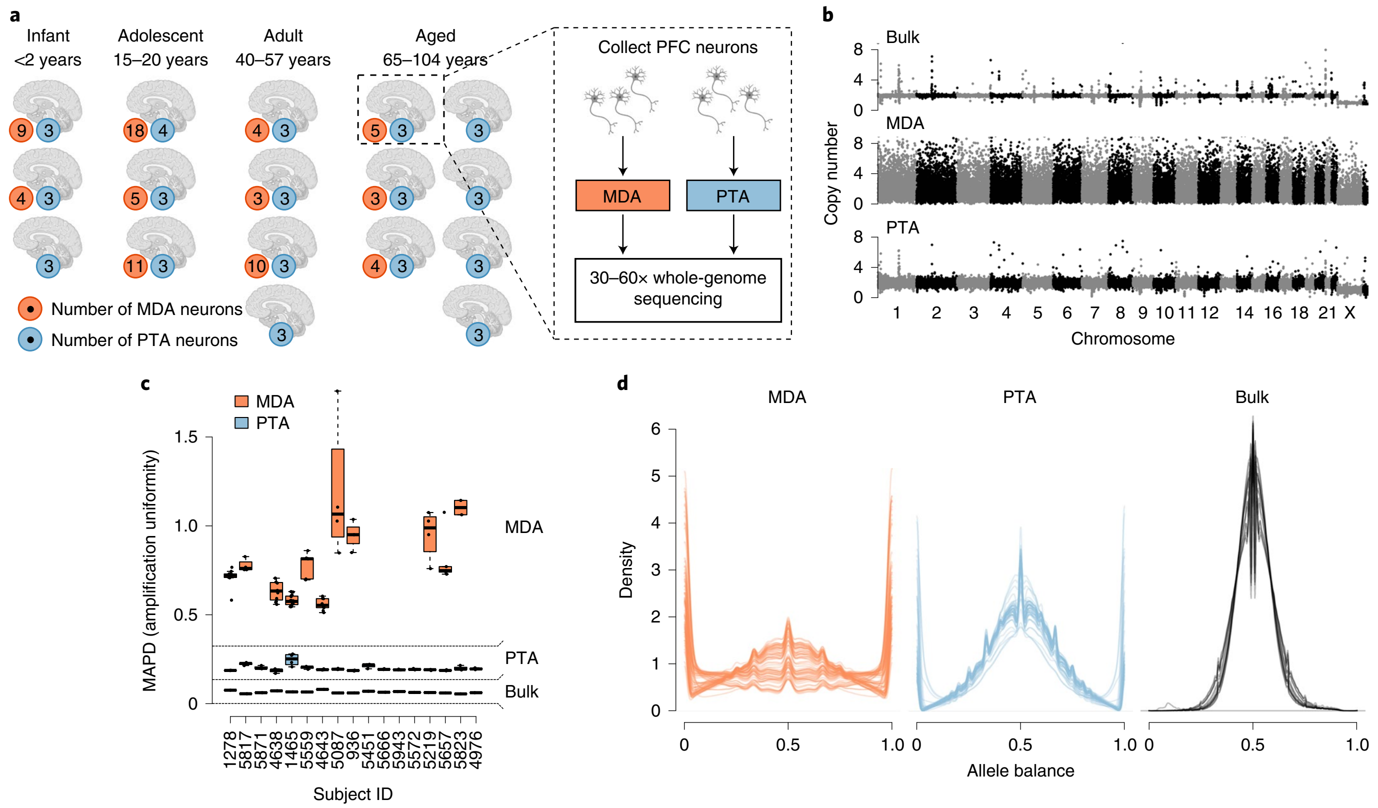


Fig. 1 | Improved large-scale amplification characteristics of PTA compared with MDA. a, Study design. Single neurons were collected from the PFCs of brains of 17 individuals ranging in age from infancy to elderly. Single neurons were amplified by either PTA or MDA and then sequenced to high coverage. Image created using BioRender.com. **b**, Representative copy number profiles for bulk (top), MDA-amplified (middle) and PTA-amplified (bottom) genomes. **c**, MAPD for MDA-amplified and PTA-amplified neuronal genomes from the same individuals. Lower values indicate better performance. The average MAPDs of MDA (0.75) and PTA (0.21) correspond to an average fluctuation in read depth between neighboring 50-kb windows of 68% and 14%, respectively. The boxplot whiskers represent the furthest outlier ≤1.5× the interquartile range (IQR) from the box, the box the 25th and 75th percentiles and the center bar the median (*n* = 17 bulk samples, *n* = 52 PTA neurons, *n* = 75 MDA neurons). **d**, Allele balance for germline heterozygous SNPs measuring the evenness of amplification between homologous alleles in a diploid cell. Each line corresponds to one single cell or bulk sample. Values near 0.5 indicate balanced amplification of homologous alleles; values near 0 or 1 indicate complete dropout of one allele. On average, 71% of each PTA genome was balanced (allele balance between 0.3 and 0.7) compared with only 39% of each MDA genome.

order of magnitude more false positives (FPs) than there are mutations in some nonneoplastic cells (~0.9 FPs per megabase[10]). We therefore developed SCAN2 (Single Cell ANalysis 2), a genotyper that augments the SCAN-SNV model of allelic imbalance with a new mutation signature[17] approach to increase sSNV detection sensitivity and extend analysis to somatic indels (sIndels). Applied to PTA data, SCAN2 detects somatic SNVs in scDNA-seq data with ~60-fold fewer FPs per megabase than conventional GATK calling and >5-fold fewer FPs than other single-cell SNV genotypers. Importantly, unlike phylogenetic or population genetics-based genotypers[18,19], SCAN2 is not fundamentally limited to detecting shared mutations and can thus recover nonclonal private mutations such as those that occur in postmitotic cells. Using SCAN2 and PTA, we produced a catalog of 20,090 somatic SNVs and 2,714 somatic indels from 52 healthy human neurons. Our catalog confirms a previously discovered age-related SNV signature[6] (with a slightly revised rate of accumulation) and reveals an enrichment of somatic mutations—particularly indels—in transcribed genes and brain-specific regulatory elements.

Results

PTA improves amplification quality and reduces artifacts. Using PTA, we amplified the genomes of 52 single neurons from the PFCs of 12 neurotypical individuals and sequenced to 30–60×, including 15 neurons from 5 neurotypical individuals from another study[20] (Fig. 1a and Supplementary Table 1). From 11 of the 17 individuals, 75 single neurons were previously amplified by MDA[6], providing a direct comparison between the two protocols. Despite being sequenced to lower depth, PTA-amplified neurons showed several favorable characteristics compared with MDA-amplified cells, including substantial reduction in coverage variability and allelic dropout across the genome (Fig. 1b–d). Regions of allelic imbalance were generally not reproduced between PTA amplifications, with the exception of neurons from a single subject (4638) (Extended Data Fig. 1). Surprisingly, large-scale somatic copy number mutations (>5 Mb; Methods) were detected in only 2 of the 52 PTA neurons (Supplementary Note and Supplementary Fig. 1), in contrast to the previous reports of pervasive copy number alterations in human neurons, especially in young individuals[21,22].

Amplification also creates artifactual SNVs (of the order of 10,000 per MDA amplification[23]) and indels (insertions and deletions), typically by spontaneous DNA damage or polymerase errors. Most artifacts occur late in the amplification reaction and, as a result, are not present on all sequencing reads derived from one haplotype. This leads to improper read phasing with nearby SNPs and inconsistent variant allele fractions (VAFs), enabling genotypers such as LiRA and SCAN-SNV to filter most late artifacts. Early artifacts, especially those that occur before amplification (for example,

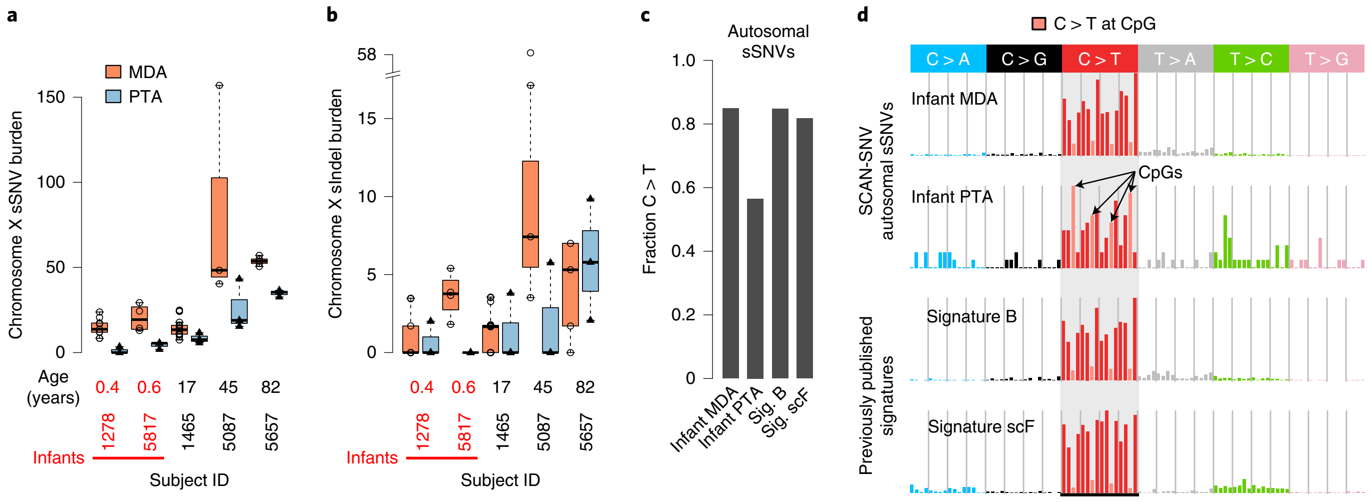


Fig. 2 | PTA identifies MDA-induced artifacts. a,**b**, Sensitivity-adjusted sSNV (**a**) and sIndel (somatic indel) (**b**) burdens per X chromosome for five male subjects with both MDA- and PTA-amplified neurons. The boxplot whiskers represent the furthest outlier ≤1.5× the IQR from the box, the box the 25th and 75th percentiles and the center bar the median ($n = 16$ PTA neurons and $n = 39$ MDA neurons). **c**, Fraction of C > T mutations among SCAN-SNV sSNV calls in infant neurons and two previously published signatures. Sig., Signature. **d**, Mutation spectra of SCAN-SNV sSNVs across 13 MDA infant neurons, 6 PTA infant neurons, the C > T-rich signature B reported by Lodato et al.[6] and the MDA artifact signature scF reported by Petljak et al.[24]. Light-red bars denote C > T mutations that occur at CpG sites.

during cell lysis), can be more difficult to identify because they are present on a larger fraction of reads. The most severe case, which we previously described[15] and refer to as single-strand dropout (SSD), occurs when no sequencing reads from the pre-artifact haplotype are present.

Haploid male X chromosomes provide an opportunity to measure the rate of SSD artifacts: as both true mutations and SSD artifacts should have near-100% VAF, a systematic excess of near-100% VAF putative mutations in MDA compared with PTA neurons from the same individual would imply the presence of SSD artifacts. Using a simple genotyping approach (Methods), we found a median excess of 15 somatic SNVs and 3.7 somatic indels per MDA X chromosome (Fig. 2a,b and Supplementary Fig. 2), corresponding to about 550 SNV and 136 indel SSD artifacts per MDA-amplified genome.

Analysis of autosomes, which includes both SSD and other MDA artifacts, identified a C > T-dominated MDA artifact signature. We focused on infant neurons, which contain the fewest age-related mutations and thus are expected to contain the highest proportion of MDA artifacts. sSNVs from MDA-amplified infant neurons were ~10-fold more abundant (282 versus 26 sSNVs per neuron) compared with those from PTA-amplified neurons despite similar detection sensitivity, enriched for C > T mutations (85% versus 59%; Fig. 2c) and resembled two signatures previously reported to be associated with technical artifacts (signature B[6] and signature scF[24]; Fig. 2d). Analysis by LiRA produced similar patterns (cosine similarity 0.988). Although PTA sSNVs were also primarily C > T, preference for CpG contexts (similar to COSMIC SBS1) suggests true somatic mutation acquisition during developmental mitoses rather than an artifactual origin. Nevertheless, raw mutation counts indicate a much lower burden of SNV artifacts compared with MDA.

SCAN2: detecting somatic SNVs and indels in PTA single cells. SCAN2 builds on SCAN-SNV, a single-cell somatic SNV genotyper that creates a genome-wide, position-specific model of allelic amplification imbalance by integrating local allele balance information indicated by the VAFs of heterozygous germline SNPs. Inspired by the characteristic mutation signature of SNV artifacts in MDA, SCAN2 incorporates signature analysis as a new source of information to further identify mutations that could not otherwise be confidently distinguished from artifacts by VAF alone. The approach operates in two passes (Fig. 3a and Methods). First, the signature of true mutations is learned by 'VAF-based' calling (which determines whether candidate mutation VAFs are consistent with local estimates of allele imbalance) with stringent calling thresholds. If individual cells do not provide a sufficient number of mutations to estimate the true signature, several single cells subject to the same mutational processes (SCAN2 provides a test of this assumption; see Methods) can be combined. Second, the newly learned true mutation spectrum is compared against a universal PTA artifact signature that we have identified (Supplementary Note and Supplementary Fig. 3) and candidate mutations rejected in the first pass may be rescued if they are unlikely to have originated from the artifact signature (Fig. 3b; see Supplementary Fig. 4 for examples of signatured-based artifact likelihood estimation).

SCAN2 performance was assessed using both simulated data (synthetic diploid X chromosomes; see Methods) and a kindred single-cell system. Varying mutation burden levels were used in both assessments because it can strongly influence the false discovery rate (FDR). For high mutation burdens (for example, germline variant detection), a genotyper's FDR may appear low because true variants (annotated SNPs and individual-specific variants in germline analysis) greatly outnumber artifacts; however, the same genotyper may produce unacceptable FDRs when artifacts outnumber mutations, as is the case at the low mutation burdens typical of healthy human cells (for example, 0.1–1.0 sSNV per Mb[5,6,9,10]; Supplementary Note and Supplementary Fig. 5).

On simulated sSNVs, SCAN2 outperformed SCAN-SNV by increasing sensitivity by ~82% (46% versus 25%) while maintaining a similar FDR (8.6% versus 9.5%) (Extended Data Fig. 2a,b). SCAN2 also outperformed two other single-cell SNV genotypers (Monovar[18] and SCcaller[23]) by several-fold reduction of the FDR (Extended Data Fig. 2c,d). For SCAN2's signature-based rescue, near-maximal performance was achieved when 500–1,000 mutations were available for learning the mutation signature of true sSNVs (Extended Data Fig. 2e,f) and SCAN2's increased sensitivity was maintained for sSNV simulations using various COSMIC signatures (range of cosine similarity to the PTA SNV artifact signature: 0.06–0.871; Extended Data Fig. 2g–i).

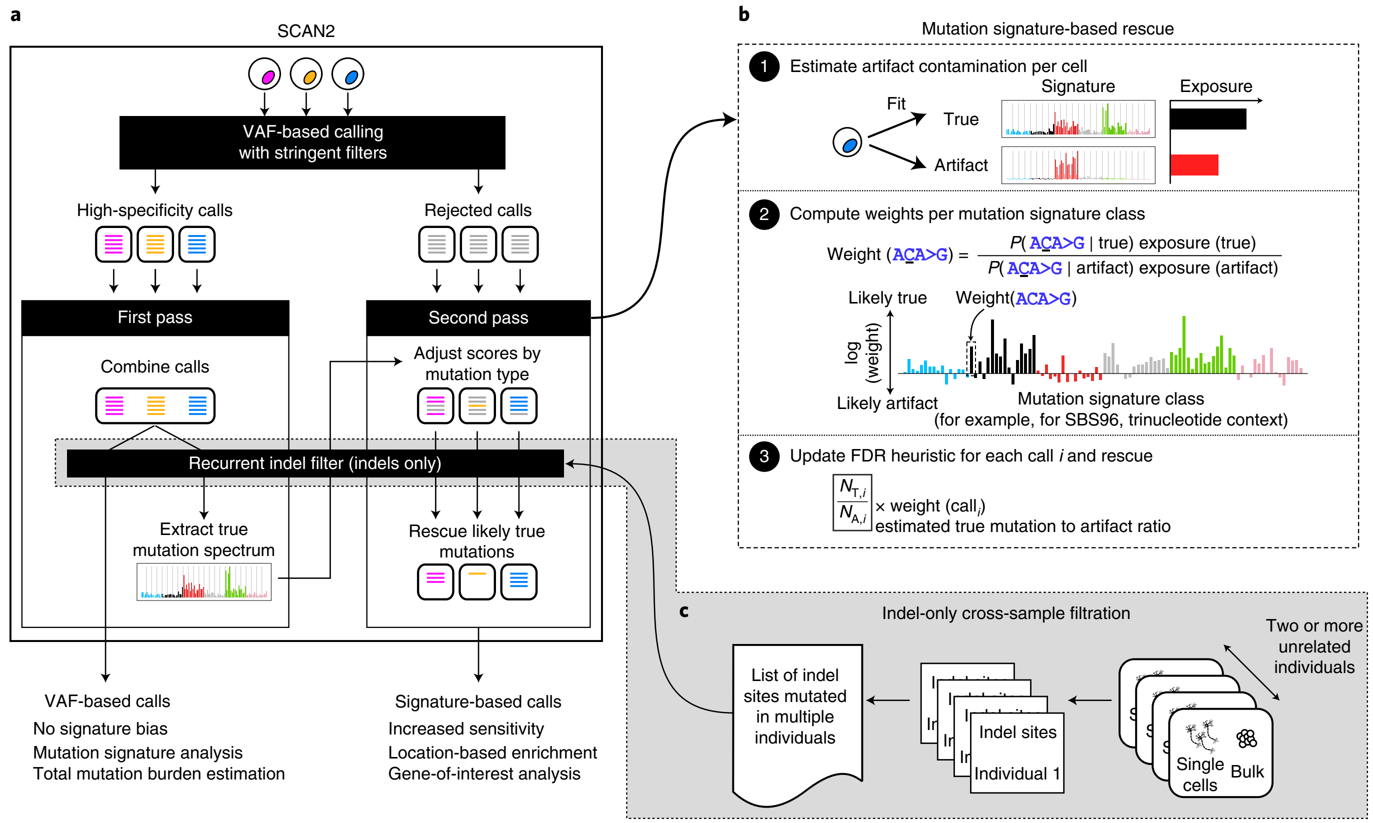

a
SCAN2
VAF-based calling with stringent filters
High-specificity calls
Rejected calls
First pass
Second pass
Combine calls
Adjust scores by mutation type
Recurrent indel filter (indels only)
Extract true mutation spectrum
Rescue likely true mutations
VAF-based calls
No signature bias
Mutation signature analysis
Total mutation burden estimation
Signature-based calls
Increased sensitivity
Location-based enrichment
Gene-of-interest analysis
b
Mutation signature-based rescue
1 Estimate artifact contamination per cell
Signature
Exposure
Fit
True
Artifact
2 Compute weights per mutation signature class
Weight (ACA>G) = P(ACA>G | true) exposure (true) / P(ACA>G | artifact) exposure (artifact)
Likely true
Weight(ACA>G)
log (weight)
Likely artifact
Mutation signature class
(for example, for SBS96, trinucleotide context)
3 Update FDR heuristic for each call i and rescue
N_T,i / N_A,i × weight (call_i)
estimated true mutation to artifact ratio
c
Indel-only cross-sample filtration
List of indel sites mutated in multiple individuals
Indel sites
Individual 1
Two or more unrelated individuals
Single cells
Bulk

Fig. 3 | SCAN2 mutation signature-based calling approach for somatic SNVs and indels. Overview of SCAN2 workflow using somatic SNV spectra for demonstration; 83-channel indel spectra are used for somatic indel analysis. **a**, SCAN2's two-pass mutation signature-based calling, in which mutation signatures from high-specificity calls are used to rescue likely true mutations from the rejected call set. Mutations may be combined across cells exposed to the same mutation processes to increase the number of VAF-based calls available when extracting the true mutation signature. This may not be necessary for cells with very high mutation burden. **b**, Candidate somatic mutations are rescored separately for each single cell given the true mutation signature learned. The likelihood of being generated by the true signature is computed for each mutation class (96-dimensional 'SBS96' for SNVs and 83-dimensional 'ID83' for indels). This likelihood acts as a prior for a previously described heuristic that estimates the number of true mutations ($N_{T,i}$) and artifacts ($N_{A,i}$), with characteristics similar to mutation candidate *i*[14]. **c**, For indel calling only, recurrent artifacts are further removed by a cross-sample list of sites where indels are observed across cells from multiple unrelated individuals.

Kindred single-cell systems further confirmed SCAN2's low FDR. In typical kindred cell analyses, somatic mutations called in one kindred cell are validated if they are also present in other kindred cells or bulk sequencing of the kindred clone. However, some true somatic mutations are private and would not be validated by this approach, resulting in an overestimated FDR. We therefore used crossbred mouse embryonic stem cell (mESC) lines, which have greatly increased SNP density (approximately tenfold greater than human SNP rates), to enable LiRA analysis across more of the genome and provide an alternative mutation validation metric. Two mESC clones were created and one was treated with aristolochic acid I (AAI) to induce a high burden of sSNVs with a known signature (SBS22) (Methods). We sequenced four PTA-amplified single cells and one clonal bulk from each clone for performance assessment. On the untreated clone, SCAN2 recovered 23% more sSNVs than SCAN-SNV (32% versus 26%) with an FDR between 9% and 32% depending on how FPs were defined (Methods and Extended Data Fig. 3a,b). SCAN2 recovered 28% more sSNVs than SCAN-SNV on the AAI-treated clone (52% versus 41%) and clearly recovered the aristolochic acid signature (Extended Data Fig. 3c). On AAI-treated cells, both SCAN2 and SCAN-SNV achieved an FDR ≈ 1%, which is expected due to the high sSNV rate induced by AAI.

A major advance in SCAN2 is the ability to identify somatic indels from scDNA-seq data. Indel detection uses a modified sSNV pipeline, offering both VAF-based and signature-based calling, but depends on an additional filter to remove recurrent indel artifacts (Fig. 3c). Although it is rare for a particular sSNV artifact to occur twice in the same amplification, processes that generate artifactual indels (for example, polymerase stutter[25] and microhomology-mediated chimera formation[26]) occur more frequently in certain genomic regions and can therefore recur, leading to inflated artifact VAFs. This effect can be further exacerbated by ambiguities that cause different indels to look alike (for example, in a homopolymer such as AAAAA, a single-base deletion of any of the five As would appear identically in sequencing data). To remove these recurrent indel artifacts, SCAN2 builds a list of sites that contain indel-supporting reads in single cells from multiple individuals; somatic indel candidates overlapping these sites are rejected.

We first adapted other methods to detect indels in simulated data, but found impractically high error rates: naive application of SCAN-SNV to indels yielded 19.9% sensitivity but 61–85% of calls were FPs; GATK HaplotypeCaller with Variant Quality Score Recalibration (using criteria similar to SCAN2 to remove germline indels, see Methods) recovered 57% of indels, but with >99% FDR. Even when adding SCAN2's recurrent indel filter to GATK HaplotypeCaller, the FDR remains high at 54–90%. Only SCAN2 was able to achieve high specificity: 33.6% (16.9% using only VAF-based calls) of spike-in indels were recovered with a mean FDR < 2% (Extended Data Fig. 4a–c). In contrast to sSNVs, indel properties such as their length often affect detection

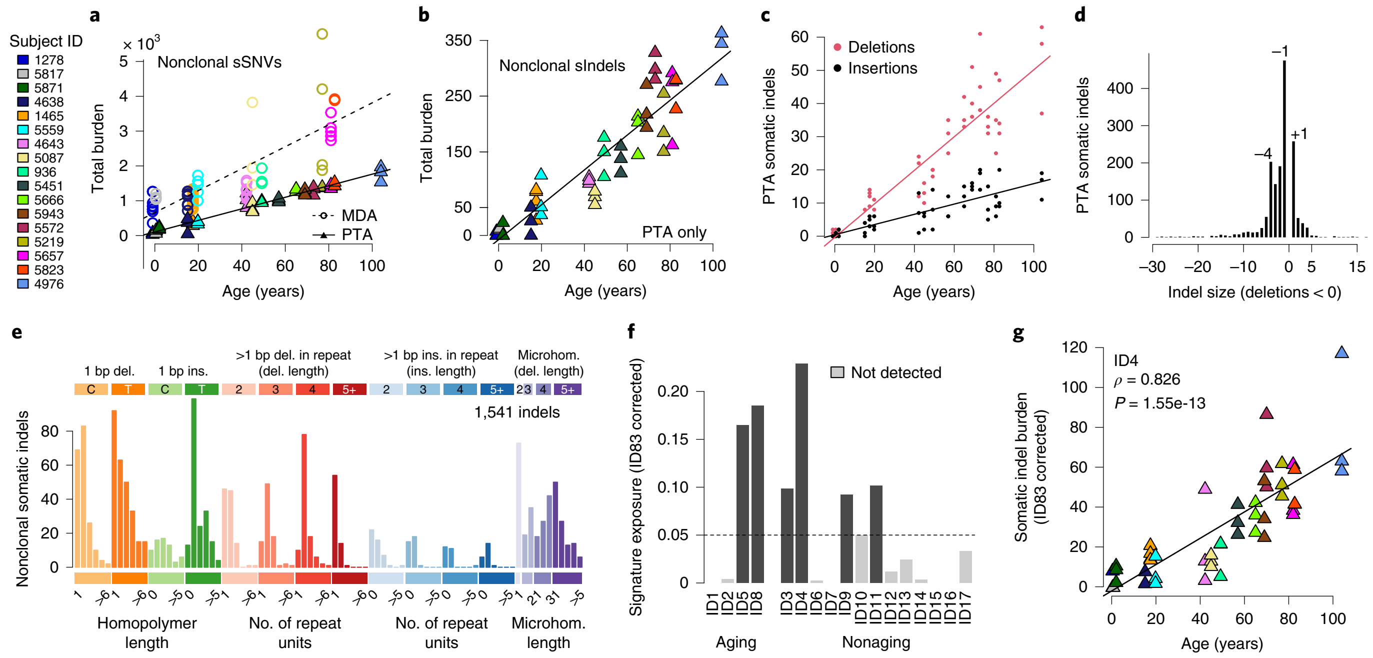


Fig. 4 | SCAN2 VAF-based somatic SNVs and indels in aging human neurons. a, Genome-wide, extrapolated accumulation rate of somatic SNVs in PTA- (triangles) and MDA- (circles) amplified single human neurons. Colors represent 17 individuals. **b**, Genome-wide extrapolated rate of somatic indel accumulation. **c**, Age-related increase of somatic insertions and deletions called from PTA neurons; raw counts are reported, not sensitivity-adjusted genome-wide rates. **d**, Distribution of somatic indel lengths from PTA neurons. **e**, Raw mutation spectrum of somatic indels. del., deletion; ins., insertion; Microhom., Microhomology. **f**, Exposures to COSMIC ID signatures calculated by least squares fitting. Exposures were corrected by normalizing indel counts by ID83 channel-specific sensitivity (Extended Data Fig. 4f) before fitting. **g**, Association of ID4, a signature of unknown etiology, with neuron age; ρ: correlation coefficient, *P* value: two-sided Student's *t*-test for correlation = 0. Trend lines in **a**–**c** and **g** are computed using mixed-effects linear regressions to account for multiple points being derived from the same individual.

sensitivity; indeed, we found reduced sensitivity for indels in homopolymers and tandem repeats of more than four units (Extended Data Fig. 4d–g).

The effects of various SCAN2 filtering steps on sSNV and indel calling are provided in Supplementary Fig. 6.

Nonclonal somatic SNV accumulation in aging human neurons. SCAN2 is also able to predict the genome-wide somatic mutation burden per cell by adjusting for somatic detection sensitivity and the fraction of the genome accessible to analysis (Methods). SCAN2 accurately predicted the number of spike-in mutations in the simulated datasets used in our performance assessment (Supplementary Fig. 7). By fitting a linear model to SCAN2 somatic SNV burden estimates from the 52 PTA-amplified neurons, we estimate that 16.5 sSNVs accumulate per year in the autosomes of human neurons (Fig. 4a). LiRA, which predicts genome-wide mutation burdens using a smaller set of very-high-confidence sSNVs, predicted a similar rate of 17 sSNVs per year, helping to validate SCAN2's approach (Extended Data Fig. 5). De novo signature analysis of VAF-based sSNVs from PTA-amplified neurons confirmed signature A, an aging-associated signature that we previously recovered from MDA-amplified neurons[6] (Supplementary Fig. 8). Notably, no signature resembling signature B was extracted from de novo analysis of our PTA sSNVs. Importantly, our filters, which require sSNVs to be undetectable in matched bulk, remove most clonal somatic mutations that occur during nervous system development. Thus, the intercept of our aging trend underestimates the somatic mutation burden at birth.

We previously estimated a yearly increase of ~23 sSNVs per year in a larger cohort of MDA neurons using LiRA[6]. Using the 74 MDA neurons in the present study, SCAN2 estimated 31 sSNVs per year in MDA neurons. De novo signature extraction recovered both signatures A and B from the combined set of MDA and PTA sSNVs. We hypothesized that, if the difference in MDA and PTA accumulation rates were due to signature B-like MDA artifacts, then its removal from MDA neurons should result in sSNV accumulation rates more consistent with PTA neurons. Indeed, after subtracting the signature B-like exposure from MDA neurons, SCAN2's yearly accumulation rate estimate decreased from 31 sSNVs per year to 22 sSNVs per year and removal of a strong elderly outlier (subject 5219) further decreased the rate to 19 sSNVs per year, more closely matching that of PTA neurons (Supplementary Note and Supplementary Fig. 9). Taken together, these observations provide compelling evidence that sSNVs accumulate in human neurons at a rate closer to 16 sSNVs per year with a signature A-like pattern, and further confirms that MDA artifacts can be largely attributed to signature B.

Characteristics of somatic indels in single human neurons. SCAN2 identified 1,541 indels from the 52 PTA-amplified neuronal genomes using VAF-based calling. Somatic indels increased with age by approximately three somatic indels per neuron per year (Methods and Fig. 4b), which is similar to rates observed in several mitotically active cell types[8–10,27]. However, our rate probably represents a lower boundary on indel accumulation owing to lower sensitivity for indels of varying length and repeat content. Deletions accumulated 3.3-fold faster than insertions (Fig. 4c) and indel sizes ranged from −29 base pairs (bp) to +17 bp (Fig. 4d). As was the case for sSNVs, MDA yielded a higher accumulation rate estimate of 6.0 somatic indels per year and we again attribute this to MDA artifacts (Supplementary Fig. 10a). Of 75 MDA neurons, 7 contained an exceptionally high number of indel calls

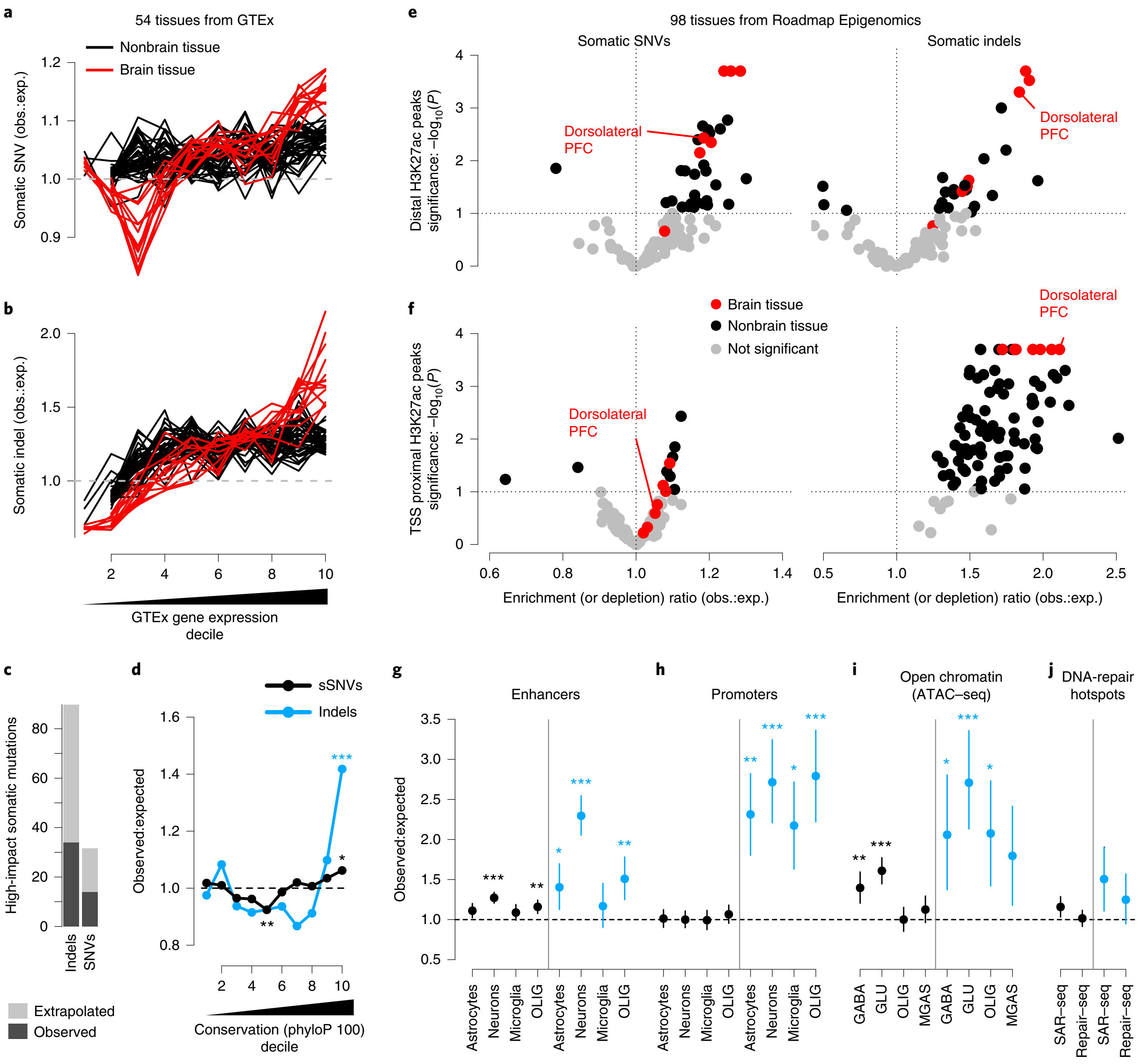


Fig. 5 | Enrichment of neuronal mutations in functionally active genomic regions with tissue- and cell-type specificity. a,**b**, sSNV (**a**) and somatic indel (**b**) enrichment compared with local gene expression levels measured by GTEx. Each line corresponds to one GTEx tissue type; tissues from primary brain specimens are shown in red. obs.:exp., observed:expected. **c**, The number of high-impact (classified HIGH by SnpEff; includes severe protein-altering effects such as stop gains, stop losses and frameshifts) sSNVs and somatic indels detected by SCAN2's signature-based approach (dark gray) and extrapolation to autosome-wide burden (light gray). **d**, Mutation enrichment compared with local sequence conservation. **e**,**f**, Enrichment analysis of neuronal mutations in H3K27ac peaks from 98 Roadmap Epigenomics tissues. H3K27ac peaks are classified according to whether they are within 2 kb of an H3K4me3 peak in the same tissue (**f**, TSS proximal) or not (**e**, distal). Distal peaks are interpreted as intergenic enhancers. Dorsolateral PFC is shown because it most closely matches the neurons sequenced in the present study. **g**–**j**, Mutation enrichment analysis of several datasets. Cell-type-specific enhancers (**g**) and promoters (**h**) are from Nott et al.[33]; cell-type-specific OCRs measured by ATAC–seq are from Hauberg et al.[34] (**i**); DNA-repair hotspots measured in induced human neurons (**j**) are reported by Wu et al.[30] (SAR–seq) and Reid et al.[29] (Repair–seq). GABA, GABAergic neurons; GLU, glutamatergic neurons; OLIG, oligodendrocytes; MGAS, microglia and astrocytes. Error bars (**g**–**j**): 95% bootstrapping CI with $n=10^4$ bootstrap samplings; center point is the observed mutation count divided by the mean mutation count over bootstrap samplings. $^*P<0.01$, $^{**}P<0.001$, $^{***}P<0.0001$ by two-sided permutation test (Methods) without multiple hypothesis correction.

characterized by single-base insertions in homopolymers of length ≥3 (Supplementary Fig. 10b–e). Due to the added artifacts, MDA indels were not included in subsequent analyses.

De novo mutation signature extraction yielded only a single spectrum (Fig. 4e) that was broadly similar to spectra from dividing cells[9,10,27], but with a greater burden of deletions (Extended Data Fig. 6a–d). Fitting the aggregate indel spectrum to the COSMIC indel catalog produced 6 indel signatures with >5% contribution; however, the COSMIC catalog is relatively new and may not contain the ID signatures relevant to neurons. Two of the four ID signatures described as clock like, ID5 and ID8 (Fig. 4f and Extended Data Fig. 6e), were detected. The absence of the two other clock-like signatures, ID1 and ID2, is consistent with the proposed etiology involving DNA replication, which cannot be active in postmitotic neurons. However, our analysis of indel sensitivity on simulated data indicated that lack of ID1 and ID2 could also be explained by low sensitivity specific to these signatures (Extended Data Fig. 4f). The most prevalent signature was ID4, a deletion-rich signature observed in several cancer types but with unknown mechanism. Surprisingly, ID4 is more strongly correlated with age in neurons than the clock-like signatures ID5 and ID8 (Fig. 4g and Extended Data Fig. 6g); correlation with age = 0.82, 0.42 and 0.69, for ID4, ID5 and ID8, respectively. ID3 was recently detected in normal bronchial epithelium[27], especially in smokers, and also shows correlation with age in neurons (correlation = 0.60). The remainder of the detected signatures (ID9 and ID11) contribute similar numbers of mutations as ID3, but are less well correlated with age.

Neuronal SNVs and indels are enriched in regulatory elements. The increased sensitivity of SCAN2's mutation signature-based approach is particularly advantageous when quantifying somatic mutation enrichment in genomic regions of interest. Using mutation signatures, SCAN2 recovered approximately 36% more somatic SNVs (20,090 versus 14,748) and 76% more somatic indels (2,714 versus 1,541) from PTA neurons compared with VAF-based calls. Only a handful of neurons showed evidence of deviation from the batch-wide sSNV and indel signatures ($P < 0.05$ for 3 out of 52 and 2 out of 52 neurons for SNV and indel signatures, respectively; statistical test described in Supplementary Note and Supplementary Fig. 11). To estimate enrichment levels in genomic regions, background mutation rates were determined by randomly permuting somatic mutations across regions of genome accessible to SCAN2 (Methods and Extended Data Fig. 7).

Spurred by reports of transcriptional strand bias in neuronal SNVs (particularly T>C mutations in ATN trinucleotide contexts[2,6]), we first compared neuronal somatic mutation density to gene expression levels from 54 tissues in the Genotype–Tissue Expression project (GTEx; Methods). In genic regions, there was a significant positive relationship between mutation burden (both sSNV and indel) and gene expression specifically for brain tissues (Fig. 5a,b), with the most expressed decile containing an ~15% increase in sSNV and a 50–100% increase in indel mutation density. Among genic mutations, there were more than twice as many high-impact (determined by SnpEff[28]) somatic indels than sSNVs, despite sSNVs outnumbering indels 8 to 1 (Fig. 5c). Indels were also strongly enriched in the 10% of the genome with the highest evolutionary conservation, with an overrepresentation of 42% (Fig. 5d).

The large number of somatic SNVs and indels identified using PTA and SCAN2 allow the analysis of both mutation types in relation to promoters[29] and promoter-distal enhancers[30], which have been recently reported to show elevated levels of DNA damage, DNA repair and double-stranded breaks in neurons[29–31]. Enhancers and promoters were defined using H3K27ac and H3K4me3 chromatin immunoprecipitation sequencing (ChIP–seq) peaks from the Roadmap Epigenomics Project (98 tissues and cell lines[32]; Methods). A significant enrichment in transcription start site (TSS)-distal enhancers was detected for both SNVs (~30% increase, ~1.3 observed:expected) and indels (~80% increase, ~1.8 observed:expected) and, critically, the most significant enrichments were seen in primary brain tissue (Fig. 5e). Near active TSSs, only somatic indels showed evidence of enrichment and it was not tissue specific (Fig. 5f). Chromatin states[32]—which offer alternative definitions of promoters and enhancers based on a combination of chromatin marks—showed similar patterns, with indel enrichment in active TSSs (ChromHMM annotation: 1_Tss) and nongenic enhancers (7_Enh; Extended Data Fig. 8). In agreement with our GTEx analysis, chromatin state analysis also revealed enrichment for SNVs and indels in weakly transcribed regions (5_TxWk), a state that often covers the bodies of transcribed genes. Strong depletion was observed for indels in inactive chromatin states such as heterochromatin (9_Het) and Polycomb repressed regions (14_ReprPcWk), whereas minor depletions were found for sSNVs in heterochromatin.

Remarkably, both sSNVs and indels showed highly significant (sSNVs and indels: $P < 10^{-4}$) enrichment in neuronal enhancers (Fig. 5g), but reduced or marginal significance in enhancers active in nonneuronal cell types (sSNVs: $P = 0.0005$, 0.017, 0.071; indels: $P = 0.0009$, 0.007, 0.255 for oligodendrocytes, astrocytes and microglia, respectively). Promoter and enhancer elements active in several brain-specific cell types were obtained from a study of FACS-purified neurons, microglia, oligodendrocytes and astrocytes[33]. Mutation enrichment levels in these cell-type-specific regulatory elements were similar to those estimated from H3K27ac peak analysis, with SNVs and indels increased by 27% and 129%, respectively. Consistent with Roadmap Epigenomics data, indels but not SNVs were enriched in promoters and did not show a preference for cell type (Fig. 5h).

Analysis of open chromatin regions (OCRs) derived from assay for transposase-accessible chromatin using sequencing (ATAC)–seq of flow-sorted γ-aminobutyric acid (GABA)-ergic and glutamatergic neurons, oligodendrocytes, microglia and astrocytes[34] provided further evidence of preferential mutation accumulation in regulatory elements. sSNVs were strongly enriched in neuron-specific OCRs whereas indel enrichment was strong but less tissue specific (Fig. 5i).

Finally, we measured enrichment of mutations in the DNA-repair hotspots recently reported by Wu et al.[30] and Reid et al.[29]. Enhancer-associated hotspots[30] were enriched for somatic indels (51% increase; 95% confidence interval (CI) 11%, 90%, $P = 0.02$) but no enrichment was found in promoter-associated hotspots[29] (Fig. 5j). sSNVs were also enriched in enhancers but with marginal significance. Notably, all enrichments presented in Fig. 5 remained robust when reanalyzed with higher minimum sequencing depth requirements, providing further evidence that local differences in sensitivity do not explain our observations (Extended Data Fig. 9).

Discussion

Our analyses of PTA-generated, single-neuron genome sequencing represent a major advance in scDNA-seq technology and provide insight into the mutagenic processes of long-lived human neurons. Direct comparison of PTA- and MDA-amplified neurons from the same brain sample identified MDA artifacts, confirmed the signature of age-related somatic SNVs and refined the estimated yearly accumulation rate of sSNVs in postmitotic human neurons. Furthermore, SCAN2 analysis of 52 PTA neurons provided mutation density profiles which, when compared against a variety of data modalities (gene expression, ChIP–seq, ATAC–seq, evolutionary conservation and coding sequence impact), provided consistent signals of mutation enrichment in functional regions of the genome. Most strikingly, the increased enrichment level of indels in brain-specific regulatory regions suggests that somatic indels may interfere with neuronal regulatory programs. For example, DNA breaks in the promoters of early response genes triggered by neuronal activity[31,35] may be responsible for some of these indels and, if true, the associated indels may be especially deleterious.

Both PTA and SCAN2 were pivotal in enabling these findings. Although PTA itself is a substantial improvement over MDA, genotypers tuned for low mutation burdens remain critical for analysis of healthy cells. SCAN2's key advantages over other tools are the ability to detect somatic indels and its use of multi-sample information (for example, mutational signatures) to enhance sensitivity for non-shared sSNVs and indels genome wide. Indeed, compared with LiRA and SCAN-SNV in this cohort of 52 PTA neurons, SCAN2 recovered 533% and 36% more sSNVs, respectively, and is the only tool designed to detect indels. For optimal SCAN2 performance, cells combined for mutation signature-based rescue should be subject to the same mutational processes (SCAN2 provides a statistical test to help discover strong violations of this assumption) and, for some analyses (for example, de novo mutation signature extraction or fitting), it may be more appropriate to use SCAN2's VAF-based calls to avoid signature-related biases. When analyzing somatic mutation density in small genomic regions (for example, within promoter or enhancer regions), we recommend correcting for local differences in nucleotide content to account for signature-related biases in SCAN2 calls, as done in the present study using permutations.

The rates and signatures of SNV and indel mutations that we report are in line with results from two recent studies using orthogonal technologies. META-CS, a single-cell amplification technique that tags Watson and Crick strands, reported an increase of ~16 sSNVs per year in neurons[36]. NanoSeq, a single-molecule consensus sequencing method for bulk DNA, estimated 17.1 sSNVs and 2.5 indels per year[37]. Our study additionally provides unprecedented power to analyze the distribution of somatic mutations in human neurons by detecting sixfold more sSNVs than the META-CS study (~20,000 versus ~3,000) and approximately fourfold more sSNVs and indels than the NanoSeq study (~20,000 sSNVs and ~2,700 indels versus ~5,000 sSNVs and 600 indels for the present study and NanoSeq, respectively). Furthermore, most of the human genome is accessible to PTA whereas other technologies can be more limited (restriction enzyme-based NanoSeq is limited to ~29% of the genome[37]). This difference in genome coverage may explain discrepancies in findings: for example, the NanoSeq study found only an association between indel burden—not sSNV burden—and transcription levels and a weak sSNV enrichment rather than depletion in heterochromatic regions.

Our study establishes a methodology for somatic mutation detection from scDNA-seq of PTA-amplified whole genomes. In particular, our approach can analyze genomes with low mutation burden and in cases where somatic mutations may not be shared by multiple cells. We anticipate that our methodology will enable a wide range of studies, including somatic mutation analysis of neurons from individuals with neurodegenerative diseases, further characterization of mutations caused by exposures to mutagenic compounds and measuring the efficiency and accuracy of CRISPR (clustered regularly interspaced short palindromic repeats) editing at the single-cell level.

Online content

Any methods, additional references, Nature Research reporting summaries, 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/s41588-022-01180-2.

Received: 23 November 2020; Accepted: 4 August 2022;
Published online: 26 September 2022

References

1. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 43–51 (2013).
2. Lodato, M. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
3. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
4. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
5. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
6. Lodato, M. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
7. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
8. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
9. Franco, I. et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat. Commun.* **9**, 800 (2018).
10. Franco, I. et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biol.* **20**, 285 (2019).
11. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
12. Evrony, G., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. *eLife* **5**, e12966 (2016).
13. Zhang, C. Z. et al. Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.* **6**, 6822 (2015).
14. Luquette, L. J. et al. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. *Nat. Commun.* **10**, 3908 (2019).
15. Bohrson, C. et al. Linked-read analysis identifies mutations in single-cell DNA sequencing data. *Nat. Genet.* **51**, 749–754 (2019).
16. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl Acad. Sci. USA* **118**, e2024176118 (2021).
17. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
18. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant detection in single cells. *Nat. Methods* **13**, 505–507 (2016).
19. Singer, J., Kuipers, J., Jahn, K. & Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nat. Commun.* **9**, 5144 (2018).
20. Miller, M. B. et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022).
21. McConnell, M. J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
22. Chronister, W. D. et al. Neurons with complex karyotypes are rare in aged human neocortex. *Cell Rep.* **26**, 825–835 (2019).
23. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
24. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294 (2019).
25. Gymrek, M. PCR-free library preparation greatly reduces stutter noise at short tandem repeats. Preprint at *bioRxiv* https://doi.org/10.1101/043448 (2016).
26. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* https://doi.org/10.1186/1472-6750-7-19 (2007).
27. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
28. 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* **6**, 80–92 (2012).
29. Reid, D. et al. Incorporation of a nucleoside analog maps genome repair sites in postmitotic human neurons. *Science* **372**, 91–94 (2021).
30. Wu, W. et al. Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440–444 (2021).
31. Madabhushi, R. et al. Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161**, 1592–1605 (2015).
32. Roadmap Epigenomics Consortium, Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
33. Nott et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
34. Hauberg, M. et al. Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* **11**, 5581 (2020).
35. Alt, F. W. & Schwer, B. DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair* **71**, 158–163 (2018).
36. Xing, D. et al. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc. Natl Acad. Sci. USA* **118**, e2013106118 (2021).
37. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Human tissue, case selection and ethical approval. Postmortem frozen human tissues were obtained from the National Institutes of Health (NIH) Neurobiobank at the University of Maryland Brain and Tissue Bank (UMBTB). Tissue collection and distribution for research and publication were conducted according to protocols approved by the University of Maryland Institutional Review Board (IRB; for UMBTB: no. 00042077) and after provision of written authorization and informed consent. Research on these de-identified specimens and data was performed at Boston Children's Hospital with approval from the Committee on Clinical Investigation (no. S07-02-0087 with waiver of authorization, exempt category 4) and processed according to an IRB-approved protocol at Boston Children's Hospital. Consent was obtained by the NIH Neurobiobank. Nondisease neurotypical individuals had no clinical history of neurological disease and were selected to represent a range of ages from infancy to older adulthood.

Isolation of single neuronal nuclei for scWGS. Single neuronal nuclei were isolated using fluorescence-activated nuclear sorting (FANS) for NeuN (neuronal nuclear protein), as described previously[6,38]. Briefly, nuclei were prepared from unfixed frozen human brain tissue, previously stored at −80 °C, in a Dounce homogenizer using a chilled tissue lysis buffer (10 mM Tris-HCl, 0.32 M sucrose, 3 mM $Mg(OAc)_2$, 5 mM $CaCl_2$, 0.1 mM EDTA, 1 mM dithiothreitol (DTT), 0.1% Triton X-100, pH 8) on ice. Tissue lysates were carefully layered on top of a sucrose cushion buffer (1.8 M sucrose 3 mM $Mg(OAc)_2$, 10 mM Tris-HCl, 1 mM DTT, pH 8) and ultracentrifuged for 1 h at 30,000*g*. Nuclear pellets were incubated and resuspended in ice-cold phosphate-buffered saline supplemented with 3 mM $MgCl_2$, filtered (40-µm pore size), then stained with Alexa Fluor-488-conjugated anti-NeuN antibody (Millipore., catalog no. MAB377X). Large neuronal nuclei were then subjected to FANS, one nucleus per well into 96-well plates.

Single-nucleus, whole-genome amplification by PTA. Isolated single neuronal nuclei were lysed and their genomes amplified using PTA, a recently developed method that pairs an isothermal DNA polymerase with a termination base[16]. PTA reactions were performed using the ResolveDNA EA Whole Genome Amplification Kit (formerly SkrybAmp EA WGA kit; BioSkryb), following the manufacturer's protocol. Briefly, single nuclei were sorted into wells containing 3 µl of Cell Buffer pre-chilled on ice, alkaline lysed on ice with MS Mix, mixed at 1,400 r.p.m. and then neutralized with SN1 buffer. SDX buffer was then added to the neutralized nuclei followed by a brief incubation at room temperature. Reaction-Enzyme Mix was added, then the amplification reaction was carried out for 10 h at 30 °C, followed by enzyme inactivation at 65 °C for 3 min. Amplified DNA was cleaned up using AMPure beads and yield determined by the picogreen method (Quant-iT dsDNA Assay Kit, Thermo Fisher Scientific). Samples were subjected to quality control by multiplex PCR for four random genomic loci as previously described[6] and by Bioanalyzer for fragment size distribution. Amplified genomes demonstrating positive amplification for all four loci were then prepared for Illumina sequencing. Most of the PTA single-cell whole-genome sequencing (scWGS) neuron experiments described in the present study were performed specifically for this report, and they are supplemented with experiments from older individuals described elsewhere[20], as indicated in Supplementary Table 1.

Library preparation for scWGS. Libraries were made following a modified KAPA HyperPlus Library Preparation protocol provided in the ResolveDNA EA Whole Genome Amplification protocol. Briefly, end-repair and A-tailing were performed for 500 ng of amplified DNA. Adapter ligation was then performed using the SeqCap Adapter Kit (Roche, catalog no. 07141548001). Ligated DNA was cleaned up using AMPure beads and amplified through an on-bead PCR amplification. Amplified libraries were selected for 300- to 600-bp size using AMPure beads. Libraries were subjected to quality control using picogreen and Tapestation HS D1000 Screen Tape (Agilent, catalog no. PN 5067-5584) before sequencing. Single-cell genome libraries were sequenced on the Illumina NovaSeq platform (150 bp paired-end) at 30× except for subjects 1278 (HiSeq, 60×) and 1465 (NovaSeq, 60×). Illumina reads were aligned to the human reference with decoy sequence GRCh37d5 (hs37d5) using BWA-MEM.

Kindred mESC clones. Pluripotent mESCs on a C57BL/6J × SPRET/Ei F1 background were grown on feeders and maintained in N2B27 medium supplemented with the glycogen synthase kinase-3 inhibitor, CHIR99021 (Axon Medchem, catalog no. 1386, 3 µM), the MEK/ERK inhibitor PD0325901 (Axon Medchem, catalog no. 1408, 0.4 µM) and mouse leukemia inhibitory factor (LIF) at 1,000 U ml^{-1}, referred to as 2i + LIF medium. The mESCs were either treated or not treated with AAI 50 µM (Sigma-Aldrich, catalog no. A5512) for 48 h, and subsequently disaggregated into single cells and plated at limiting dilution. Single-cell clones were picked after 1 week, allowed to expand for another week to provide enough DNA for bulk sequencing and single cells were sorted for PTA. Single cells, clones and the initial mESC line were sequenced to 30× on the Illumina NovaSeq platform (150 bp paired-end) and aligned to GRCm38 using BWA-MEM.

Single-cell amplification quality metrics. Median absolute pairwise differences (MAPDs) were computed by estimating copy number (CN) in bins CN_i of size 50 kb following ref. [39]; subsequently, MAPD = median ($|\log_2(CN_i) - \log_2(CN_{i+1})|$). CN profiles in Fig. 1b were produced using Ginkgo[40] with variable bin size 100 kb and pseudoautosomal regions masked. Allele balance distributions were computed for each neuron by rounding single-cell VAFs to three decimal places at all heterozygous SNP sites used to train the SCAN2 allele balance model and then applying R's `density` function.

Genome-wide allelic imbalance analysis. Phased training human hSNPs for each cell (located in path/to/SCAN2_output/ab_model/[single_cell]/hsnps.tab) were mapped to 1-kb nonoverlapping tiles across autosomes from GRCh37d5. The allele balance for tile i containing hSNPs $\{j\}$ is $\mathbf{A}_i = \sum_j H_{j,1} / \sum_j (H_{j,1} + H_{j,2})$, where $H_{j,k}$ is the number of reads at hSNP j supporting haplotype k. The heatmap in Extended Data Fig. 1e was produced by `pheatmap` with default parameters on the correlation matrix of $\mathbf{A}$ vectors.

Comparison of MDA and PTA somatic mutation calls. Both MDA- and PTA-amplified neurons were available for five male subjects. For X-chromosome analysis, GATK HaplotypeCaller (v.3.8.1) was run in joint mode across all samples (bulk, PTA and MDA) for each individual using the Single Nucleotide Polymorphism Database (dbSNP) 147_b37_common_all_20160601 and parameters `--dontUseSoftClippedBases -rf BadCigar -mmq60`. Genomic VCF (GVCF) joint calling was not used because information can be lost compared with providing all BAMs to the same instance of HaplotypeCaller. Pseudoautosomal regions were excluded. The resulting VCF was filtered for mutations using GATK SelectVariants `-selectType SNP -selectType INDEL -restrictAllelesTo BIALLELIC -env -trimAlternates`. Somatic SNVs and indels in single cells were called separately using the following criteria: VAF > 90%, single cell depth >median(single-cell depth), 0 alternative reads in bulk, bulk depth >10 and absence from dbSNP. A set of germline SNPs and indels for estimating sensitivity was defined by sites with bulk VAF > 90%, bulk depth >median(bulk depth) and no more than two reference reads in bulk. For each single cell, somatic sensitivity was approximated as the fraction of these germline sites passing the somatic filters (except 0 alternative reads in bulk and absence from dbSNP). The final estimated number of mutations was calculated by (no. of corrected calls) = (no. of somatic mutations called)/(estimated sensitivity).

For the autosomal sSNV comparison in infant neurons, SCAN-SNV commit 5905707 was run on all MDA, PTA and bulk data for subjects 1278 and 5817 separately (Supplementary Table 2). SCAN-SNV was run with `--target-fdr=0.01` and the same external data as in SCAN2 analysis of single human neurons.

Somatic indel detection with SCAN-SNV. To adapt SCAN-SNV for indel calling, SCAN-SNV commit 5905707 was first run (with the same calling parameters and data resources as SCAN2) to fit the AB model for each synthetic diploid (SD). Somatic indel candidate loci were identified by requiring a sum of ≥2 mutation-supporting reads across the 63 SDs, single-cell read depth ≥10, bulk depth ≥10, 0 mutation-supporting reads and a 0/0 GATK genotype string in the matched synthetic bulk. Loci present in dbSNP v.147_common were further excluded. Local AB at each somatic indel candidate was estimated by SCAN-SNV's `infer.gp` function with `chunk=1` and `flank=1e5`. All SCAN-SNV statistical tests and filters for sSNVs were applied to indel candidates with a target FDR of 0.01.

Somatic indel detection with GATK HaplotypeCaller. GATK HaplotypeCaller was run jointly on all SDs and the matched synthetic bulk with the same parameters as in section 'Comparison of MDA and PTA somatic mutation calls'. For each SD, an indel VCF was created by running GATK SelectVariants with `-selectType INDEL -select 'vc.isBiallelic()' -env -trimAlternates` and removing any indel with a nocall (./.) in either the synthetic bulk or SD being analyzed. GATK VQSR (Variant Quality Score Recalibration) was then run using the recommended parameters: VariantRecalibrator was first run with `–mode INDEL –maxGaussians 4 -resource:mills,known=false,training=true,truth=true, prior=12 Mills_and_1000G_gold_standard.indels.b37.vcf -resource:dbsnp,known=true,training=false,truth=false, prior=2 dbsnp_147_b37_common_all_20160601.vcf` followed by ApplyRecalibration with `–mode INDEL --ts_filter_level 90.0`. To remove germline and clonal mutations, candidate indels must be supported by 0 reads in bulk and >2 reads in the single cell, >10 reference bulk reads and ≥10 total reads in the single cell, and must not be present in dbSNP.

Synthetic diploid X-chromosome simulations. SD X chromosomes[14] were used to assess the performance of SCAN2 and other callers. SDs are created by merging chromosome-X reads from two male single cells (or matched bulks) from different subjects. This recreates allelic amplification imbalance and preserves real amplification artifacts. Nine SDs with 30× mean depth were generated by making all pairings of the 3 PTA cells from subjects 1278 and 5817 and downsampling the reads in each BAM to ~15×. The youngest subjects (aged 0.4 and 0.6 year) were chosen to minimize the number of endogenous somatic mutations. Endogenous mutations were identified by applying GATK HaplotypeCaller v.3.8 jointly with the nine SDs, six original PTA BAMs and two matched bulks using the same parameters as in 'Comparison of MDA and PTA somatic mutation calls'. An additional HaplotypeCaller run with `-mmq 1` was also performed. Sites satisfying the following filters in the original, full-depth PTA BAMs were considered endogenous somatic mutations: VAF $\geq$ 90% and <2 reference reads, depth $\geq$5 in the single cell, depth >10 in the matched bulk and no mutation-supporting reads in bulk in either the mapping quality 60 or mapping quality 1 runs. A single cluster of sSNVs, identified by these filters at chrX:77471371-77471423, caused by clipped alignment, was manually excluded. No endogenous indels were identified.

Each SD received a burden of 10, 25, 50, 100, 250, 500 and 1,000 SNV and indel spike-ins, for a total of 63 SDs. Random spike-in positions were uniformly sampled from chromosome X excluding assembly gaps (https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gap.txt.gz), 5-bp windows centered on each nonreference site reported by GATK in subject 1278 or 5817 and 5-bp windows centered on all sites in dbSNP v.147 common. Somatic SNV spike-ins following COSMIC signatures SBS1, SBS11, SBS12, SBS16, SBS19, SBS2, SBS23, SBS3, SBS30, SBS32, SBS4, SBS5, SBS54, SBS6, SBS7b, SBS88 and SBS9 were created by generating batches of SNVs and downsampling to match the signature being simulated. This process was iterated until the desired number of spike-ins was generated. SDs with COSMIC signatures were created only with burden = 1,000 SNVs. Somatic indel spike-in candidates further required random lengths; candidates were generated and classified (by first left-aligning indels by `bcftools norm` and then using `SigProfilerMatrixGenerator`[41] to determine ID83 status) until >1,000 candidates were obtained for each ID83 class. Somatic indel spike-ins were further required to be >150 bp from the nearest indel spike-in candidate to prevent crowding in repetitive tracts. SNV and indel spike-ins were not allowed to overlap. SCAN2 was run jointly on the set of 63 SDs, 6 full-depth PTA BAMs, 2 matched bulks and 1 synthetic bulk with the same parameters used in the analysis of single neurons. Sensitivity was calculated as the fraction of known spike-ins called; any call not in the endogenous sSNV or spike-in sets was considered an FP. Due to the ambiguous nature of indel representation, indel calls were considered matches to known spike-ins if either (1) the calls matched the spike-in indel exactly or (2) the called indel was the correct length and was located exactly 1 bp away from the spike-in location.

To better approximate real-world performance, SD candidate mutations were combined with autosomal somatic mutation candidates from single-cell 5817PFC-A before analysis with SCAN-SNV and SCAN2. This allows the N_T/N_A FDR heuristics to be computed on a full genome of data, which should better reflect real-world performance.

SNV calling with Monovar. Monovar commit `7b47571` was used and somatic SNVs were called following the authors' protocol[18]. BAMs were input to samtools mpileup v.1.9 with options `-BQ0 -d10000 -q 40`, which was piped into the monovar.py script with options `-p 0.002 -a 0.2 -t 0.05 -m 2` as recommended by the authors. To determine whether SNVs were somatic or germline, samtools was run with the same options on matched bulk data. Somatic SNVs were determined by the following filters: Monovar's single-cell genotype must not match `./.` or `0/0`; single-cell depth $\geq$10 with at least 3 mutation-supporting reads; bulk depth $\geq$6 and $\leq$1 mutation-supporting read; and single-cell VAF $\geq$ 10% for sSNVs with >100 depth or VAF $\geq$ 15% for sSNVs with depth between 20 and 100. Finally, sSNVs were filtered if any other call occurred within 10 bp.

SNV calling with SCcaller. SCcaller v.1.1 was run following the authors' recommendations. BAMs were converted to pileups using samtools v.1.3.1 with the option -C50 and hSNPs were defined using dbSNP v.147 common. Single-cell sSNVs were called by applying SCcaller's `-a varcall`, `-a cutoff` and reasoning v.1.0 script in sequence with default parameters. As recommended on SCcaller's Github README, passing somatic mutations were required to have VAF > 1/8, filter status = `PASS`, bulk status = `refgenotype` and must not have been observed in dbSNP. The standard calling parameter is $\alpha = 0.05$, whereas the stringent calling parameter is $\alpha = 0.01$.

SNV calling with LiRA. LiRA v.`1f4cab4` was run following instructions on Github. The joint VCF produced internally by SCAN2 (path/to/scan2/gatk/hc_raw.mmq60.vcf) for each individual was supplied as the input VCF to LiRA. All samples were processed as male to restrict calls to the autosomes and use a single genome size for burden estimation. Current LiRA versions used a genome size of $G = 6.349$ for males, so LiRA burden estimates were multiplied by 5.845/6.349 to match the autosomal extrapolation presented here and in ref. [6]. LiRA burden estimates retrieved from Supplementary Table 5 of ref. [6] did not require this correction.

Kindred mESC analysis. LiRA v.`3bc0ae1` was used with the global option `reference_identifier GRCm38` and the `--force` flag to `lira varcall` following the authors' instructions on Github. SCAN2 commit d8edd85 was configured with `scan2 config --target-fdr 0.01 --callable-regions True --gatk gatk3_joint --score-all-sites --parsimony-phasing` (Supplementary Note). SCAN2 data sources were: reference genome GRCm38, the SHAPEIT2 1000 genomes reference panel (ignored by `--parsimony-phasing`) and a customized dbSNP database of SPRET_EiJ sites from mgp.v.5.merged.snps_all.dbSNP142.vcf from https://www.sanger.ac.uk/data/mouse-genomes-project. One aim of the kindred analysis was to approximate real-world SCAN2 performance in human cells with a nonsimulated truth set. It was therefore necessary to reduce the high SNP density of crossbred mice (~33 million SNPs per genome) to avoid an overly accurate AB model. The SCAN2 pipeline was manually halted after rule training_hsnps_helper and the output files path/to/scan2/abmodel/[sample]/hsnps.{tab,vcf} containing training hSNPs were downsampled to ~2 million random sites by R's `sample` function. The SCAN2 pipeline was then restarted.

For FDR calculations using the standard kindred approach, sSNVs were considered true mutations if and only if they satisfied any of: VAF $\geq$ 20% in the kindred clone bulk; $\geq$5 reads in another kindred cell; or $\geq$1 read in $\geq$2 other kindred cells. For LiRA-based FDR, sites with UNLINKED status were removed and FDR was defined as the fraction of sites with status FILTERED_FP.

For sensitivity calculation, a truth set of clonal sSNVs was constructed separately for each clone using the following criteria: at least 10 reference reads and no mutation-supporting reads in the initial mESC population bulk; 50% $\geq$ VAF < 100% and depth $\geq$10 in the kindred clone being analyzed; and VAF = 0 in the other kindred clone. A total of 130 clonal SNVs was identified in the untreated clone and 17,002 SNVs were detected in the AAI clone. Reported sensitivities are the mean fraction of clonal sSNVs recovered across the four cells from each clone.

SCAN2 analysis of single human neurons. SCAN2 v.0.9 was run separately for each of the 17 subjects; for each subject, all MDA, PTA and bulk samples were provided to SCAN2. Nondefault parameters to SCAN2 were: `--abmodel-chunks=4 --abmodel-samples-per-chunk=5000 --target-fdr=0.01 --somatic-indels --somatic-indel-pon path/to/filter.rda`. SCAN2 data resources: human reference genome GRCh37d5, SHAPEIT2 phasing panel 1000GP_Phase3 and dbSNP v.147_b37_common_all_20160601. All following scan2 commands used SCAN2 v.1.0. The cross-sample filter (`--somatic-indel-pon`) was generated by `scan2 makepanel` with all 128 MDA and PTA single cells and 17 bulks supplied via the `--bam` flag. Mutations from all 52 PTA samples were combined and supplied to `scan2 rescue --rescue-target-fdr 0.01`. MDA calls were not included in signature-based rescue. Two neurons were excluded from analysis: MDA neuron 5087pfc-Rp3C5, due to high mutation burden (both in ref. [6] and the present study) and PTA neuron 4638-Neuron-4, due to a very low mutation burden.

Per-cell total mutational burdens were computed separately for sSNVs and indels using `mutburden.R` (SCAN2 0.9). Current versions of SCAN2 compute burdens automatically. Yearly mutation accumulation rates were derived from a mixed-effects linear model to account for subject-specific effects. Mixed-effects model fitting was performed separately for sSNVs and indels using the `lme4` (ref. [42]) R package with the command `lmer(age ~ total_burden + (1|subject))`. `total_burden` refers to the SCAN2 total burden estimate for each neuron.

De novo signature extraction was performed by `SigProfiler`[43] on VAF-based calls from PTA neurons only, which produced a single signature for both sSNVs and indels. Fits to COSMIC indel signatures used COSMIC v.3 signatures ID1–17. For the discovery of active signatures in Fig. 4f, all 1,541 VAF-based indels were combined and exposures to each of the 17 signatures were estimated by least squares (`lsqnonneg` from the `pracma` R package). For correlation of signature exposure with age, indels from each cell were kept separate. For indels, differing sensitivities among the ID83 channels were corrected before `lsqnonneg` by dividing by the channel-specific sensitivities derived from SD X chromosomes (Extended Data Fig. 4f).

Functional impact of point mutations. The severity of sSNV and indel mutations reported in Fig. 5c was derived from SnpEff[29] v.4.3t using the hg19 database. Duplicate and clustered mutations were removed as described in Enrichment analysis of somatic mutations. High-impact mutations were those annotated as HIGH in the first reported 'ANN' field. Extrapolation from called mutations to the expected number over the PTA cohort was obtained by dividing mutation counts by the cohort-wide sensitivity estimates of 48.7% for sSNVs and 46.2% for somatic indels.

Enrichment analysis of somatic mutations. To prevent regions with localized artifacts from driving functional impact or enrichment signals, duplicate mutation calls (that is, exact recurrence of a mutation) were either removed or downsampled to one call. For duplicate calls occurring in more than one subject, all instances were removed; for duplicate mutations in more than one neuron from the same subject (1.1% of sSNVs, 0% of indels), one occurrence was arbitrarily retained. An additional 57 sSNV calls were removed due to duplicate observations (not SCAN2 calls) in more than one subject with target.fdr <50%. Clustered mutations (any mutation within 50 bp of another mutation in a single neuron: 1.5% of sSNVs, 4% of indels) were also removed.

Permutation testing was used to generate the expected number of somatic mutations for enrichment analysis. Permutations with matching mutational signatures to the neuronal set were generated by `scan2 permtool` (SCAN2 v.1.0) with default parameters. For each mutation set S consisting of N_S mutations, 10,000 permutation sets P_i of size N_S mutations each were generated. The positions of permutated mutations were uniformly selected from the subset of the single neuron genome (in which the corresponding mutation in S was called) with single-cell depth >5 for sSNVs (≥10 for indels). Permutated mutations were then downsampled to match the SBS96 spectrum (or ID83 spectrum for indels) of S. This step controls for the expected signature bias of SCAN2-rescued calls and nucleotide content bias in the genomic region of interest. Enrichment over any genomic region R is the number of R-overlapping mutations in S divided by the average number of R-overlapping mutations from the 10,000 permutated datasets P_i. A two-sided P value is calculated by counting the number of permutation sets with greater absolute log(fold-change) than observed. CIs for enrichment estimates are computed by bootstrapping the observed mutation set and computing enrichment as described above 10,000×. To analyze enrichments with higher-sequencing-depth cutoffs D (Extended Data Fig. 9), mutations in S with depth $<D$ were removed and permutation locations were further restricted to the subset of each single neuron genome with depth $\geq D$.

Genomic covariates for enrichment analysis. GTEx expression values for 54 tissues were downloaded from https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz. Gene coordinates were obtained from https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_26/GRCh37_mapping/gencode.v26lift37.annotation.gtf.gz and isoforms were collapsed into a single record using https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model/collapse_annotation.py.

GRCh37d5 autosomes were tiled with 1-kb nonoverlapping windows and the average read depth across the 52 PTA cells was computed. Windows with mean depth <6 or mean depth in the top 2.5% of windows were removed. The remaining windows were assigned a genic coverage-weighted transcripts per million (TPM) value of the gene overlapping the window multiplied by the fraction of the window covered by the gene. If multiple genes overlap a region, the gene with highest expression is used. Windows that were <80% covered by genes were removed and considered to be intergenic. Finally, windows were ranked into deciles by their genic coverage-weighted TPM values and windows within each decile were merged to create ten regions.

H3K4me3 and H3K27ac narrowPeak files for the 98 epigenomes with H3K27ac data were downloaded from the Roadmap Epigenomics Project server. H3K27ac peaks were classified as TSS proximal if they occurred within 2 kb of an H3K4me3 peak from the same epigenome; otherwise they were considered to be TSS distal. ChromHMM 15-state mnemonic BED files were downloaded from the Roadmap Epigenomics Project server for 127 epigenomes. For each of the 15 ChromHMM states, a single merged region was created. Brain samples were defined as those with ANATOMY=BRAIN and Type=PrimaryTissue. The phyloP 100-way track was downloaded from the Universiity of California, San Cruz (UCSC) genome browser in BigWig format; average phyloP scores were computed over the same 1-kb tiling used for GTEx expression analysis, including removal of low- and high-depth windows, using the UCSC `bigWigAverageOverBed` v.2 program. Bins were then ranked into deciles by average phyloP score and windows within each decile were merged to create ten regions. Cell-type-specific enhancer and promoter regions[33] were extracted from Supplementary Table 5 tabs astrocyte enhancers, astrocyte promoters, and so on. Enhancer or promoter regions were merged within each cell type to produce two regions per cell type. OCRs for GABA, glutamate, oligodendrocytes and microglia and astrocytes from dorsolateral PFC[34] were downloaded from https://bendlj01.u.hpc.mssm.edu/ggoma. DNA Synthesis Associated with Repair sequencing (SAR-seq) DNA-repair hotspots[30] were downloaded from the Gene Expression Omnibus (accession no. GSE167257, GSE167257_SARseq_iNeuron_OverlapRep123.peaks.bed.gz); Repair–seq peaks[29] were obtained from Supplementary Table 1 of ref. [28].

Statistics and reproducibility. No statistical method was used to predetermine sample size. All 4 PTA neurons from the brain of subject 1465 were excluded from CN analyses, one PTA neuron from the present study (4638-Neuron-4) was excluded from the model of age-associated mutation accumulation due to a very low mutation burden. No other PTA neurons were excluded from any analysis. One MDA neuron from a previous study[6] (5087pfc-Rp3C5) was excluded from most analyses due to a high mutation burden. The experiments were not randomized and the Investigators were not blinded to allocation during experiments and outcome assessment.

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

Data availability

All MDA-amplified single neurons and matched bulks listed in Supplementary Table 2 were downloaded from dbGaP, accession no. phs001485.v1.p1. Only neurons from the PFCs of individuals for which additional PTA data were generated were used. Raw sequencing read data for PTA-amplified human neurons can be downloaded from dbGaP, accession no. phs001485.v3.p1. PTA-amplified mESC kindred cells and bulks can be downloaded from the National Center for Biotechnology Information's Sequence Read Archive, accession no. PRJNA832209.

Code availability

SCAN2 is available for download at https://github.com/parklab/SCAN2. Additional scripts used in the present study are available at https://github.com/parklab/SCAN2_PTA_paper_2022 and *Zenodo*[44].

References

38. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
39. Baslan, T. et al. Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
40. Garvin, T. et al. Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).
41. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).
42. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
43. Alexandrov, L. SigProfiler. MATLAB Central File Exchange https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler (2020).
44. Luquette, L. SCAN2_PTA_paper_2022. Zenodo https://doi.org/10.5281/zenodo.6532827 (2022).

Acknowledgements

We thank R. S. Hill, R. Mathieu and L. (Sahithi) Cheemalamarri at the Boston Children's Hospital & Harvard Stem Cell Institute Flow Cytometry Research Facility, the Research Computing group at Harvard Medical School, and the Boston Children's Hospital Intellectual and Developmental Disabilities Research Center Molecular Genetics Core for assistance. Human tissue was obtained from the NIH Neurobiobank at the University of Maryland, and we thank the donors and families for their invaluable contributions to the advancement of science. This work was supported by the Bioinformatics and Integrative Genomics training grant (no. T32HG002295 to L.J.L.), grant nos. K08 AG065502 and T32 HL007627 (to M.B.M.), the Brigham and Women's Hospital Program for Interdisciplinary Neuroscience through a gift from Lawrence and Tiina Rand (to M.B.M.), the donors of the Alzheimer's Disease Research program of the BrightFocus Foundation (no. A20201292F to M.B.M.), the Doris Duke Charitable Foundation Clinical Scientist Development Award (no. 2021183 to M.B.M.), PRMRP Discovery Award (no. W81XWH2010028 to Z.Z.), the Edward R. and Anne G. Lefler Center postdoctoral fellowship (to Z.Z.), and grant nos. R00 AG054748 (to M.A.L.), R01 AG070921 (to C.A.W.) and R01NS032457 and U01MH106883 (to P.J.P. and C.A.W.), and the Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation (to C.A.W.). C.A.W. is an investigator at the Howard Hughes Medical Institute. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

L.J.L. conceived and implemented SCAN2. C.A.W., M.B.M. and Z.Z. conceived the application of PTA to single neurons. P.J.P. and C.A.W. conceived and supervised the overall project. L.J.L. and Y.Z. performed computational analysis. L.J.L., M.B.M. and Z.Z. analyzed and interpreted results, and wrote the manuscript. All authors reviewed and edited the manuscript. M.B.M., Z.Z., J.G., S.B., S.K. and M.A.L. collected tissue specimens, isolated single neuronal nuclei and performed PTA amplification and amplification quality control studies. C.L.B. and L.J.L. performed LiRA analysis and comparisons to SCAN2. T.H. and C.L. generated mESCs. J.I.G. conceived and performed the mESC kindred experiment. C.L.B., A.G. and J.K. collected and processed all sequencing data. D.G. and H.J. made suggestions for signature analysis. C.G. and J.W. provided PTA reagents and advice on optimal use.

Competing interests

The authors declare the following competing interests: C.G. is Director and cofounder and J.W. is CEO and cofounder of Bioskryb, Inc., the manufacturer of PTA kits used in the present study. C.A.W. is a consultant for Maze Therapeutics (cash, equity), Third Rock Ventures (cash) and Flagship Pioneering (cash), none of which have any relevance to the present study. The remaining authors declare no competing interests.

Additional information

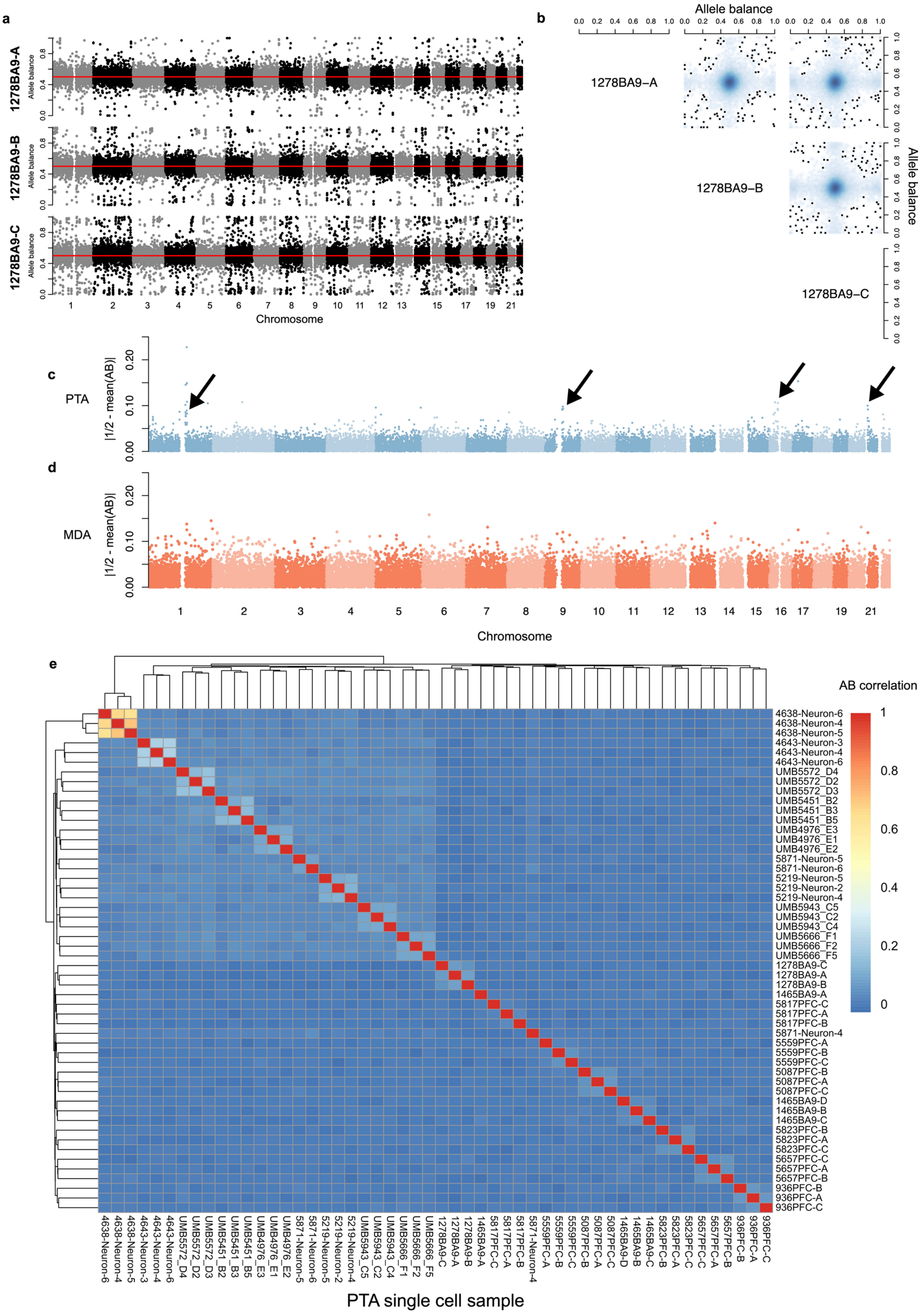
Extended data are available for this paper at https://doi.org/10.1038/s41588-022-01180-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01180-2.

Correspondence and requests for materials should be addressed to Christopher A. Walsh or Peter J. Park.

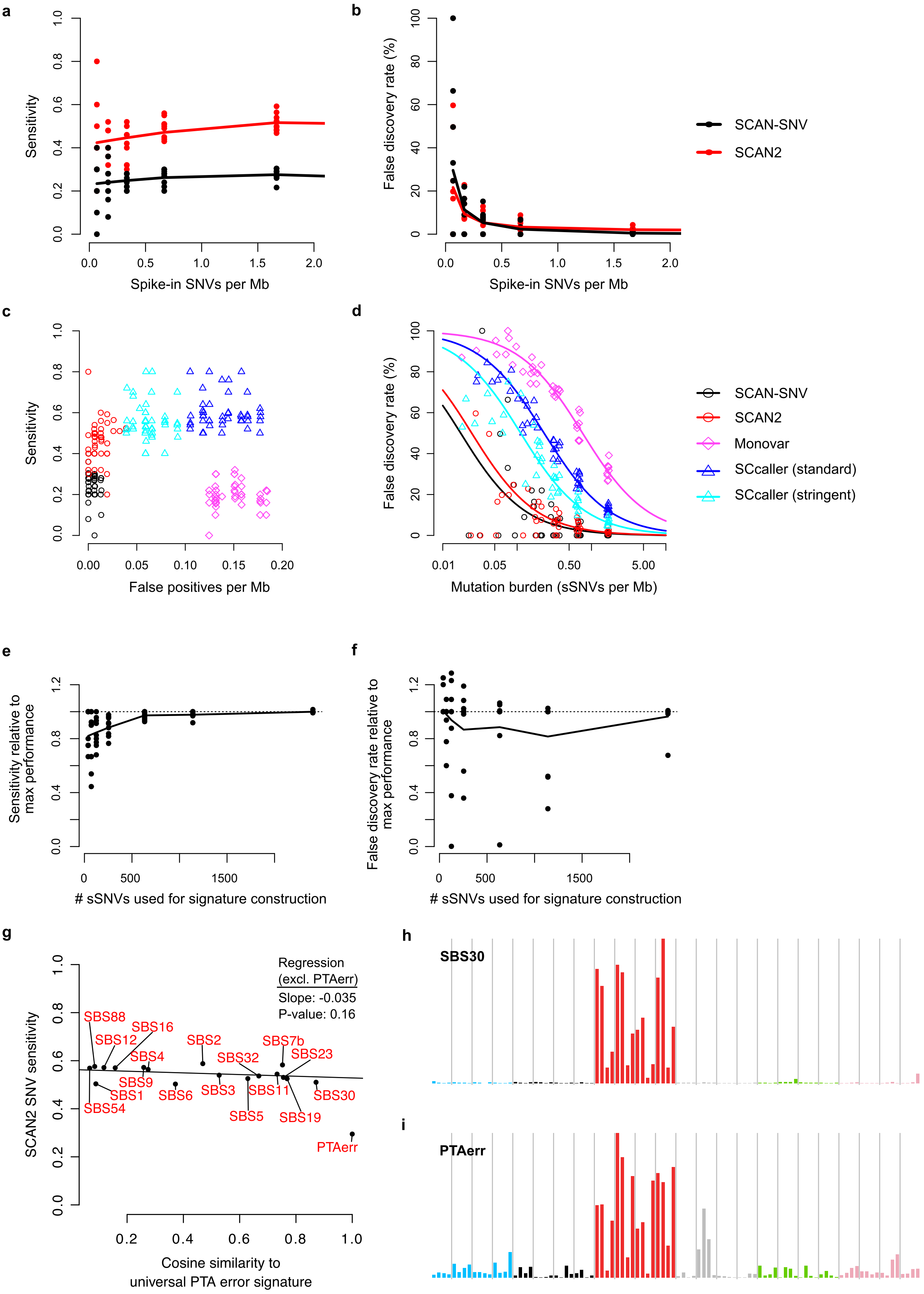
Peer review information *Nature Genetics* thanks Ruben van Boxtel and Federico Abascal for their contribution to the peer review of this work.

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



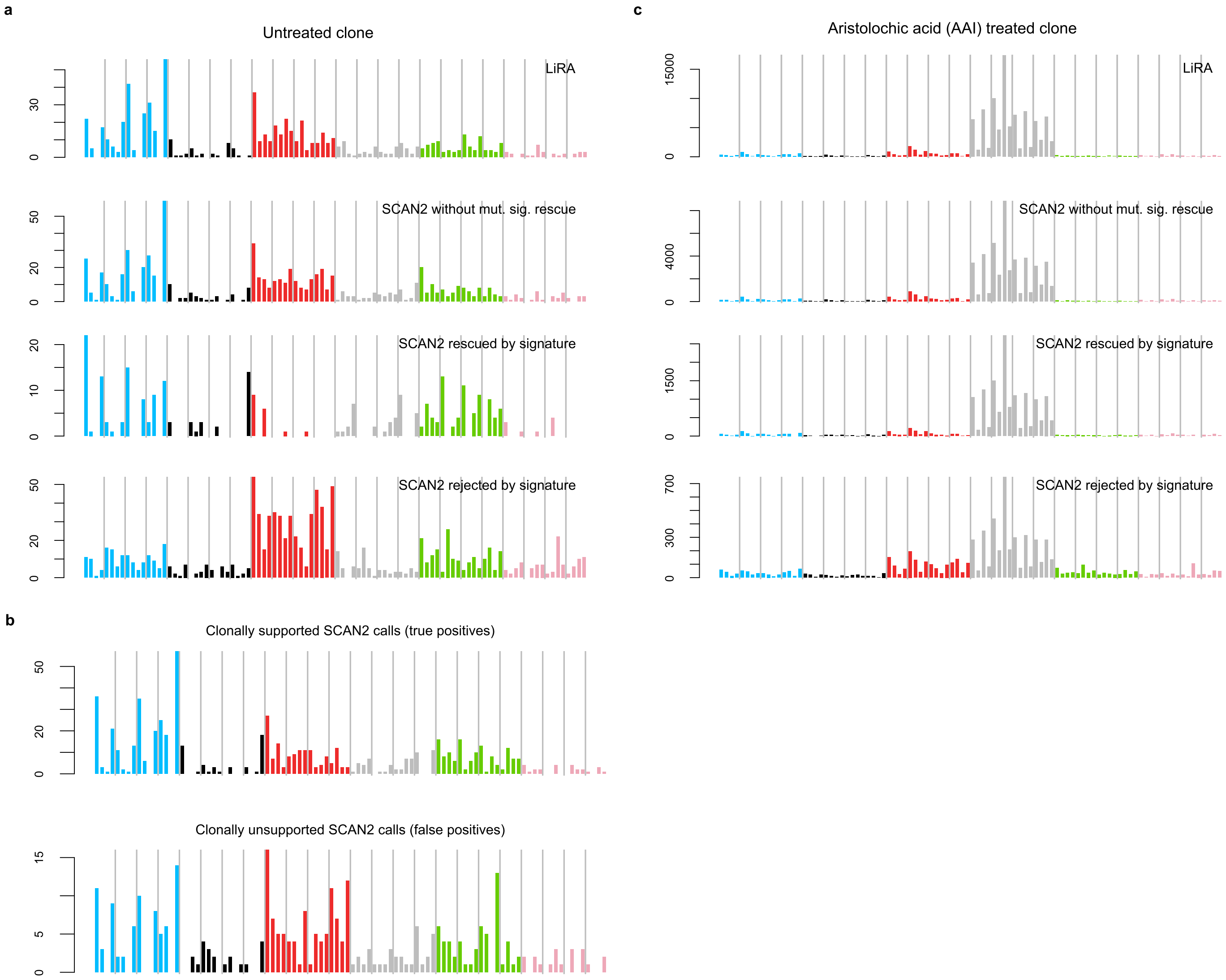
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Allele balance is not generally correlated between PTA amplifications. a. Genome-wide allele balance (binned in 100 kb windows) for 3 typical PTA cells from the same individual. **b**. Allele balance for cells in (a) plotted against each other. **c-d**. Allele balance averaged across the cohort of 52 PTA cells (c) or 75 MDA cells (d); that is, each point represents the average allele balance for a single 100 kb window. A small number of regions show consistent allelic imbalance across many amplifications (arrows). **e**. Correlation of allele balance profiles between all pairs of PTA cells. Correlation is generally low; cells from the same individual show slightly higher correlations; and a single individual (4638) shows an atypically strong correlation.

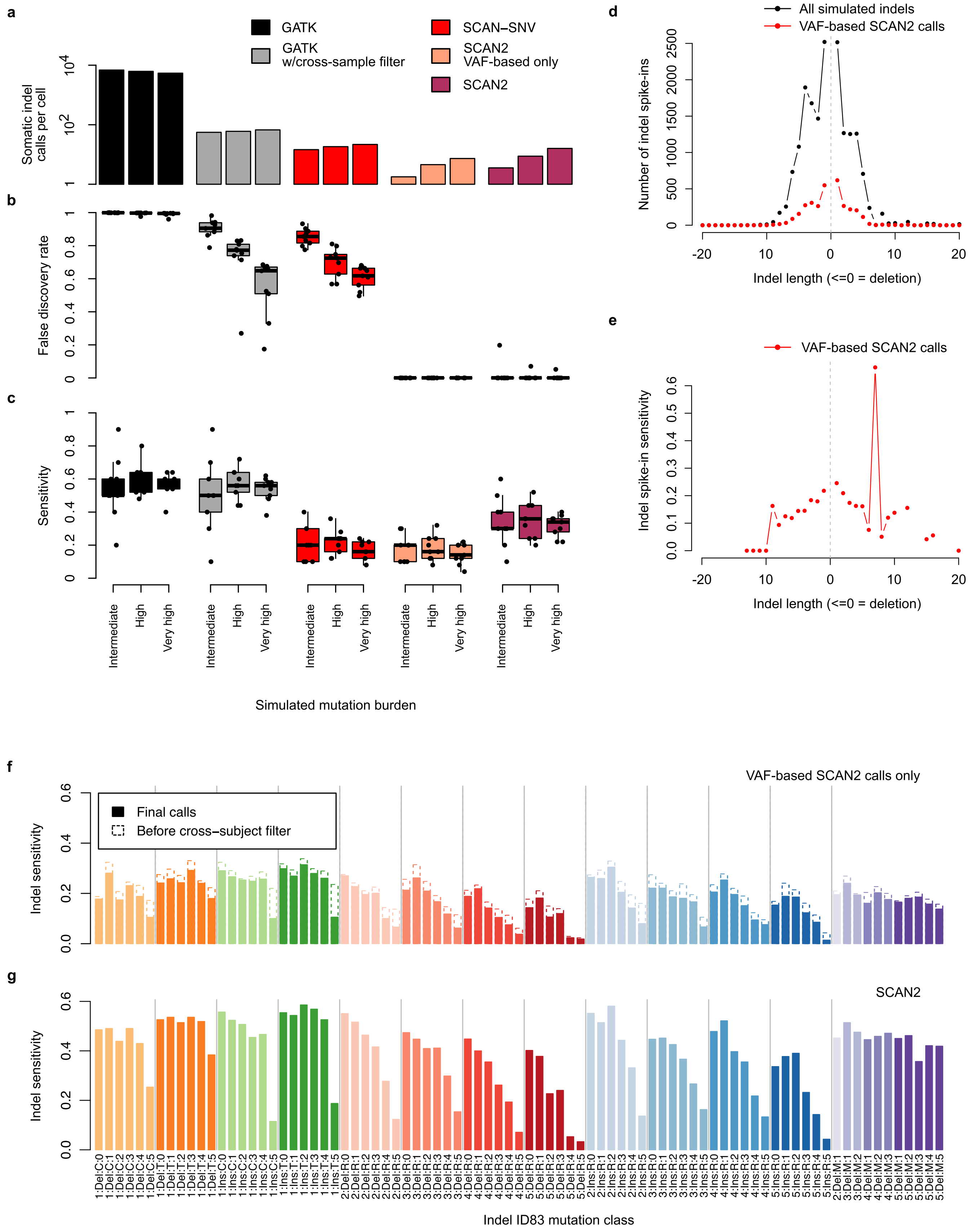


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | SCAN2 performance on simulated sSNVs. sSNVs were simulated using the synthetic diploid (SD) X chromosome approach (Methods). Sensitivity is the fraction of known spike-ins recovered and false positives (FPs) are defined as calls that are neither known spike-ins nor somatic mutations endogenous to the haploid X chromosomes used to create each SD. Each point in **a-d** represents a single SD simulation with 10-250 spike-ins. **a-b**. Comparison of SCAN2 and SCAN-SNV sensitivity (**a**; lines are R loess() fits) and false discovery rates (**b;** lines are linear regression fits to FDR ~ 1/mutations per Mb). **c-d**. Comparison to other single cell SNV genotypers. **c**. Sensitivity vs. false positives per megabase of analyzed sequence. **d**. False discovery rate vs. the number of spike-ins per megabase. Lines are parameterized by mean sensitivity *S* and false positive rate per megabase *F* measured across all points: $\text{FDR} = F / (F + xS)$. SCcaller standard uses a calling threshold of $\alpha = 0.05$ while stringent calling uses $\alpha = 0.01$. **e-f**. Performance of SCAN2 mutation signature-based rescue as a function of the number of sSNVs available for learning the true mutation signature. Sensitivity (**e**) and false discovery rate (**f**) are shown relative to the sensitivity or false discovery rate of the same SD simulation using the maximum sSNV catalog of 4,666 sSNVs. $\varepsilon = 0.0001$ was added to all quantities to avoid division by zero. Solid lines are fitted by R's loess() function. **g**. Effect of mutation signature of spike-ins on SCAN2 sensitivity. Each point is the average sensitivity of 9 SD simulations with 1000 spike-ins from a single COSMIC SBS signature. Mutation signatures are characterized by their similarity to the PTA SNV artifact signature. Solid line: linear regression on all points except PTAerr. SBS30 (**h**) is the most similar COSMIC signature to the PTA SNV artifact signature (PTAerr) (**i**).

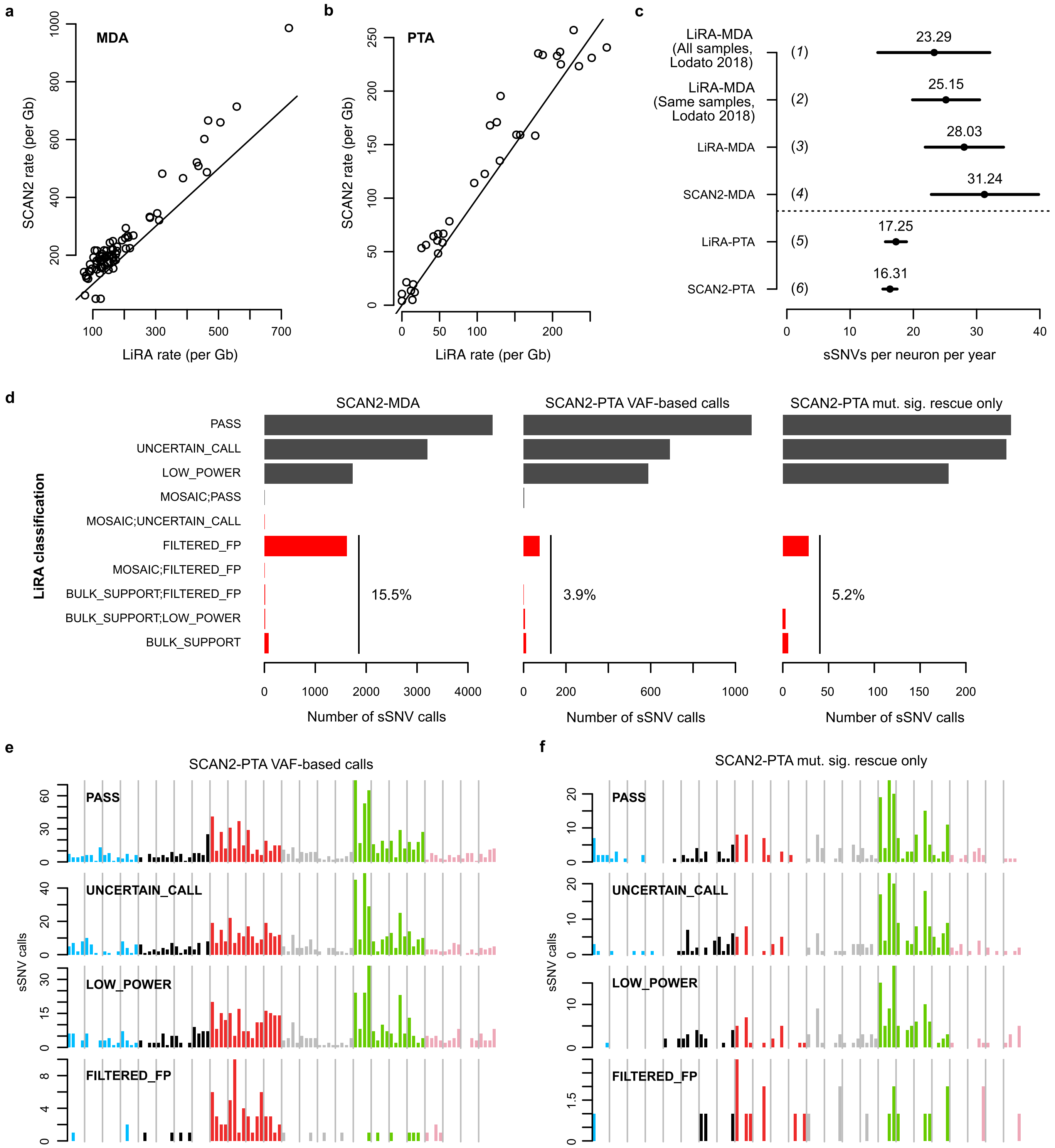


Extended Data Fig. 3 | Mutation spectra of SCAN2 and LiRA calls on kindred mouse ESC cells. a-b. SBS spectra of somatic SNVs called in 4 single cells from the untreated clone. C > A mutations (blue peaks) are characteristic of COSMIC SBS18 and the mutation signature of SNVs acquired during clonal expansion[5]. These peaks persist in the clonally unsupported SNVs (b), suggesting that the method for classifying true positives is overly conservative. **c**. Spectra for SNVs called in the 4 single cells taken from an aristolochic acid (AAI)-treated clone.



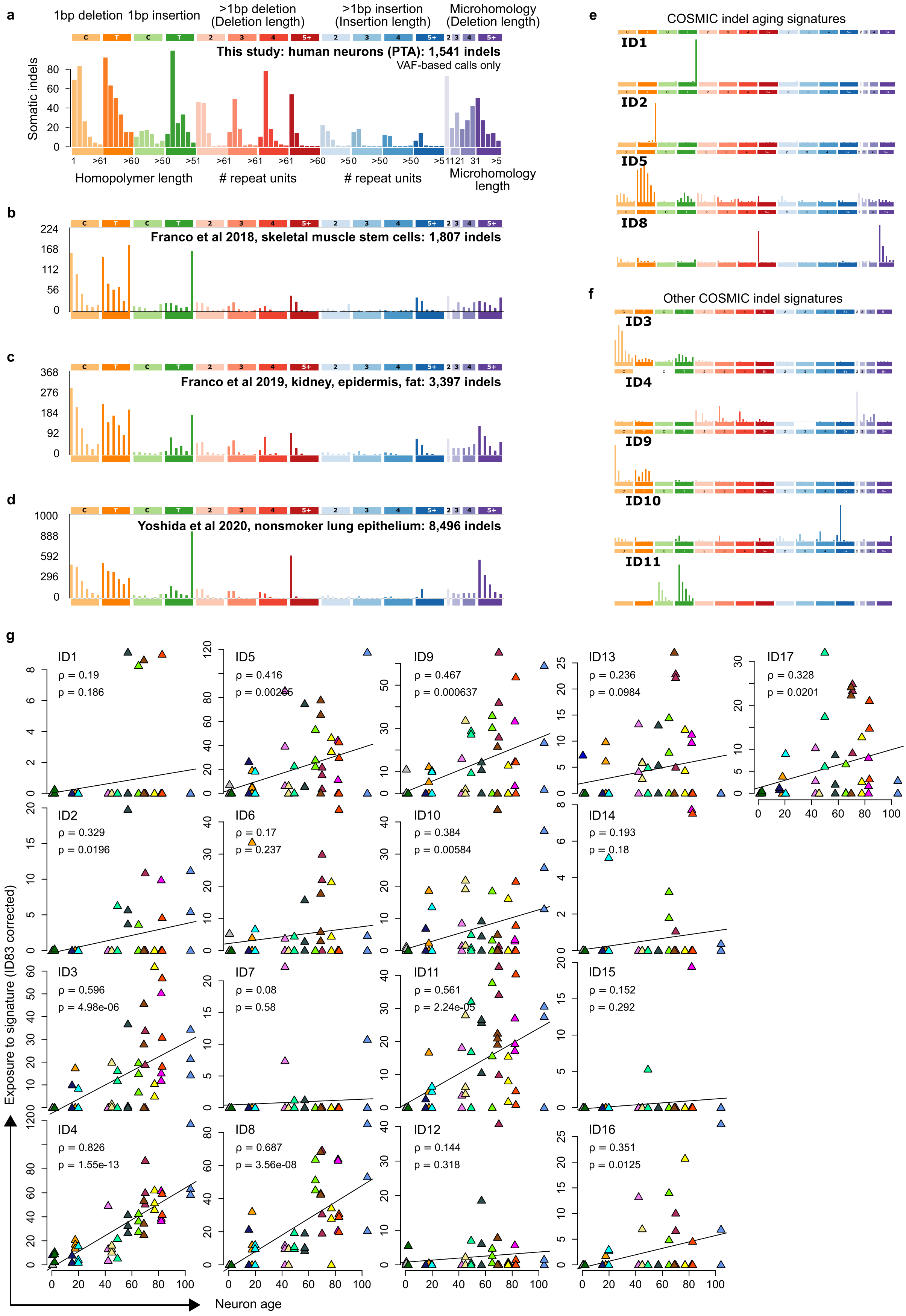
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | SCAN2 performance on simulated somatic indels. a-c. SCAN2 and other callers were applied to simulated indels using the synthetic diploid (SD) X chromosome spike-in approach (Methods). SDs received 10, 25 or 50 indel spike-ins each, which correspond, respectively, to genome-wide burdens of approximately 170 (intermediate), 430 (high) and 850 (very high) somatic indels. Performance was measured by the average number of indels called per SD (**a**), the fraction of false positives per indel call set (**b**) and the fraction of spike-ins recovered (**c**). Tested methods were SCAN2 (with and without signature-based rescue), GATK HaplotypeCaller, GATK HaplotypeCaller with filtration by SCAN2's cross-sample recurrent artifact filter and an adaptation of SCAN-SNV's somatic SNV discovery approach to indels. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median; n=9 SDs per boxplot. **d**. Distribution of indel lengths among all simulated indels (black) and VAF-based SCAN2 indel calls (red). **e**. Spike-in indel sensitivity by length for VAF-based SCAN2 calls. **f**. Sensitivity for VAF-based SCAN2 indel calling stratified by the 83-dimensional indel classification scheme used by COSMIC indel signatures (ID83). Dotted outlines: sensitivity before applying cross-subject filtration. **g**. ID83-stratified indel sensitivity for SCAN2 calls with signature-based rescue.



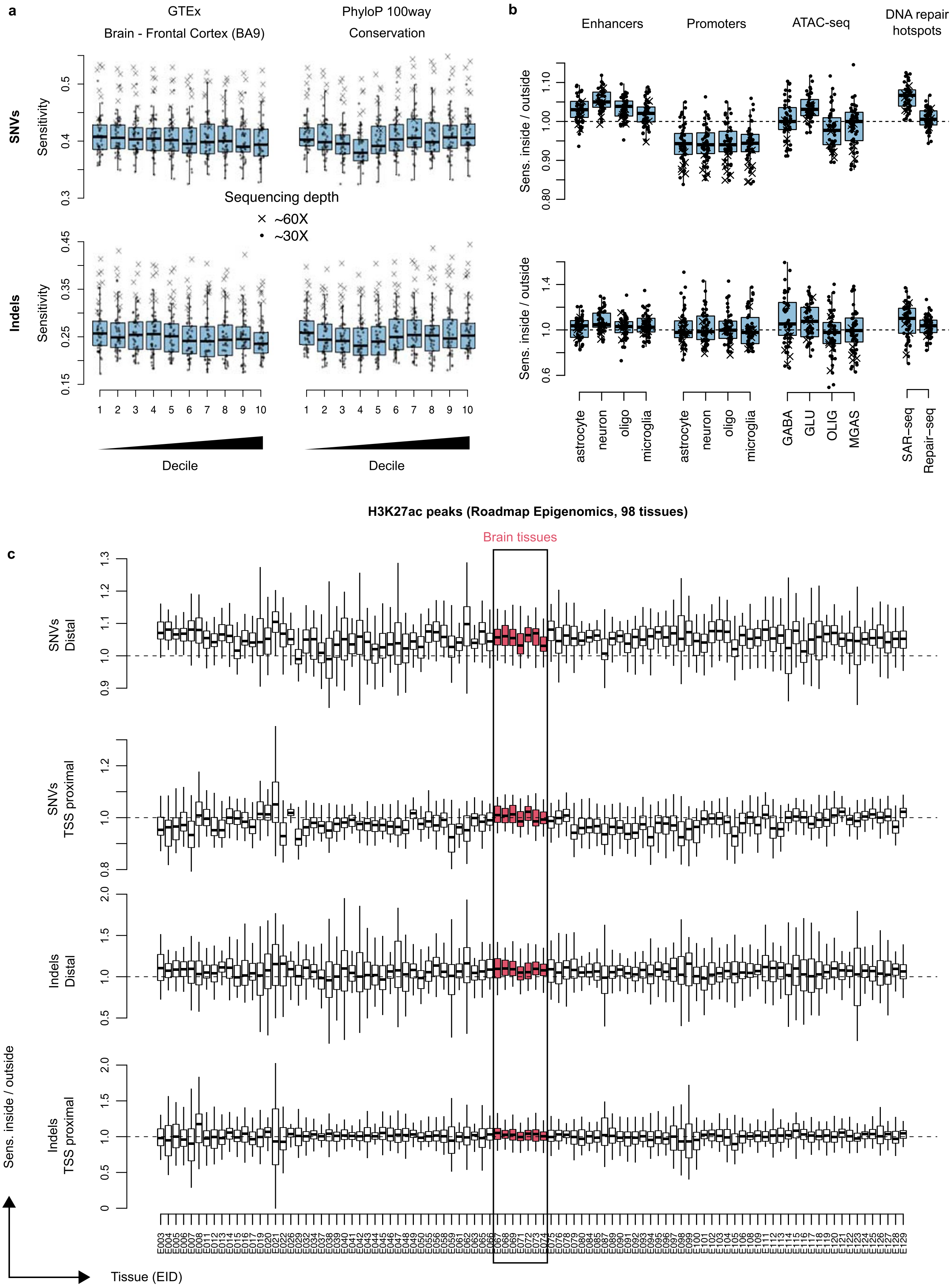
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Comparison of SCAN2 and LiRA sSNV calls on human neurons. Single human neurons were analyzed by LiRA[15], a specific but lower sensitivity approach for calling somatic SNVs. **a-b**. SCAN2 and LiRA extrapolations for the total (not called) sSNV burden per diploid Gb of human sequence from MDA- (**a**) and PTA-amplified (**b**) single neurons. Solid lines: y=x. **c**. Linear regression estimates for the number of sSNVs accumulated per neuron per year from several sources and analyses. Horizontal bars represent 95% C.I.s produced by confint applied to an lmer fit by the lme4 R package; centre points from fixef applied to the same fits. (*1*) LiRA rates taken from ref. [6], which used a larger set of *n*=91 MDA-amplified PFC neurons; (*2*) LiRA rates taken from ref. [6] using *n*=73 of the 75 MDA-amplified PFC neurons from subjects analyzed in this study (the two excluded neurons are 5087pfc-Rp3C5, an extreme outlier, and 4638-MDA-14); (*3*) rerun of LiRA on *n*=74 MDA-amplified neurons in (*2*) using the same input provided to SCAN2; (*4*) SCAN2 on *n*=74 MDA-amplified neurons; (*5*) LiRA on *n*=34 PTA-amplified neurons from donors also analyzed in ref. [6] (N.B. LiRA's higher rate estimate in (c) occurs despite lower burden estimations in (b) due to differences in model intercepts: SCAN2 intercept=95.83, LiRA intercept=17.63); (*6*) SCAN2 on all *n*=52 PTA-amplified neurons generated here. **d**. LiRA classification of SCAN2 calls where reads linked to nearby germline heterozygous SNPs are available (black: likely true sSNVs, red: possible false positives). PASS is the highest quality LiRA class. UNCERTAIN and LOW_POWER indicate lack of linking reads to make a confident call, but no evidence of artifactual status is detected. All other classes (red) are interpreted as false positives. Percentages show the fraction of all false positive classes among SCAN2 calls. **e-f**. Raw mutation spectra for SCAN2 calls without (**e**) and with mutation signature-based calling (**f**) SCAN2 calls stratified by LiRA classification. The similarities between PASS and the two lower quality UNCERTAIN_CALL and LOW_POWER classes suggest that the majority of UNCERTAIN_CALL and LOW_POWER SCAN2 calls are true mutations. Confident false positives (FILTERED_FPs) possess a C > T dominated signature with lack of C > Ts at CpGs.



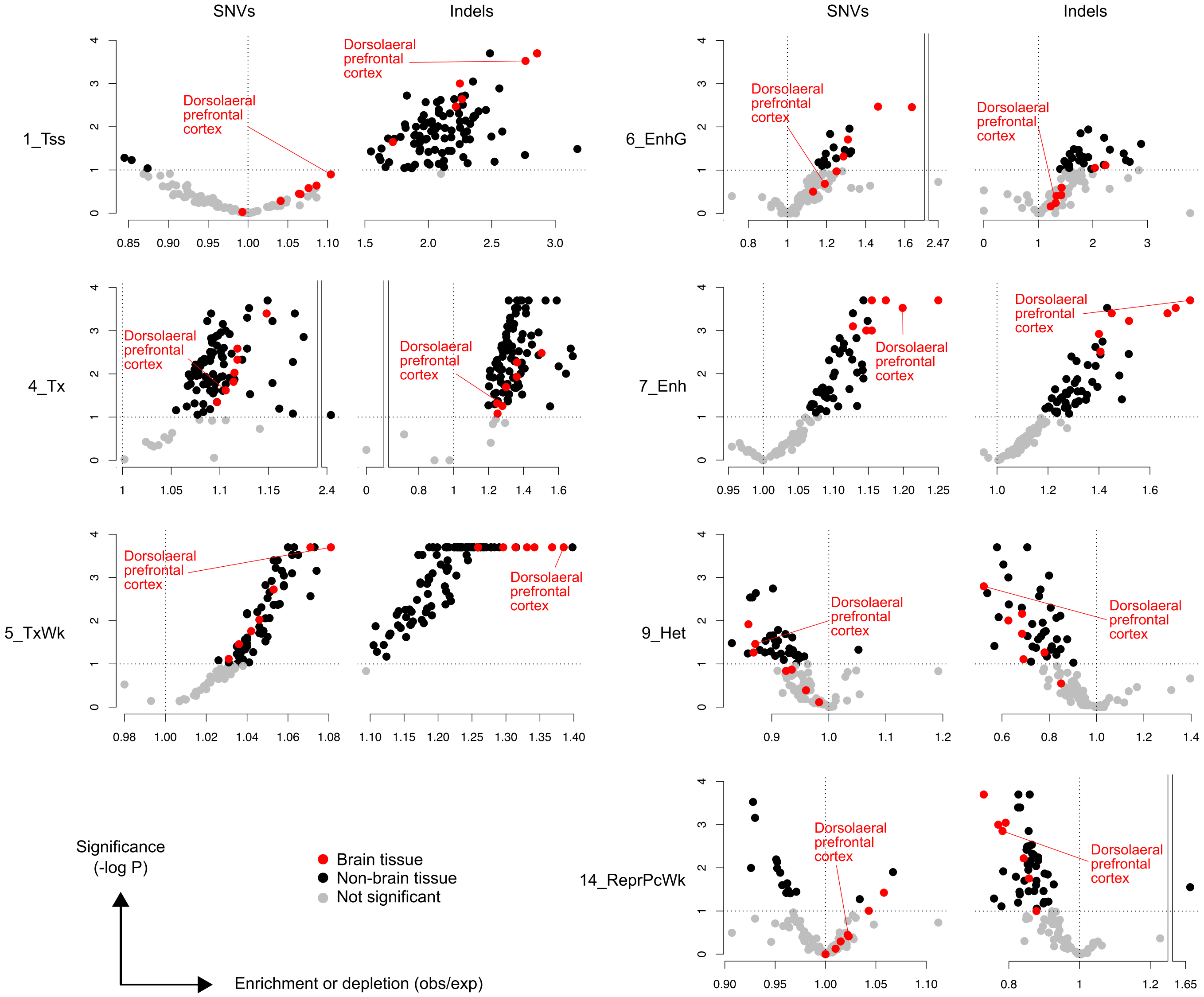
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Somatic indel mutation spectra in human neurons and other cells. a. Spectrum of 1541 indels from PTA neurons from this study, same as Fig. 4e. **b-e**. Somatic indel spectra from other studies: clonally expanded single skeletal muscle stem cells (**b**), clonally expanded single kidney (excluding hypermutated kidney cells, designated KT2 in the original study), epidermis and fat cells (**c**) and clonally expanded bronchial epithelial cells from children and never-smokers (**d**). **e**. COSMIC signatures with clock-like or age-associated annotations. **f**. Non-aging COSMIC signatures with >5% contribution to single neurons. **g**. Per-neuron COSMIC signature fits, corrected for ID83 sensitivity (Methods). Correlation (ρ) between age and exposure and *P*-value of two-sided *t*-test for correlation=0 (p) are shown for each COSMIC signature. *P*-values were not adjusted for multiple comparisons. Colors correspond to subject IDs as shown in Fig. 4. Note that y-axes are not the same scale.

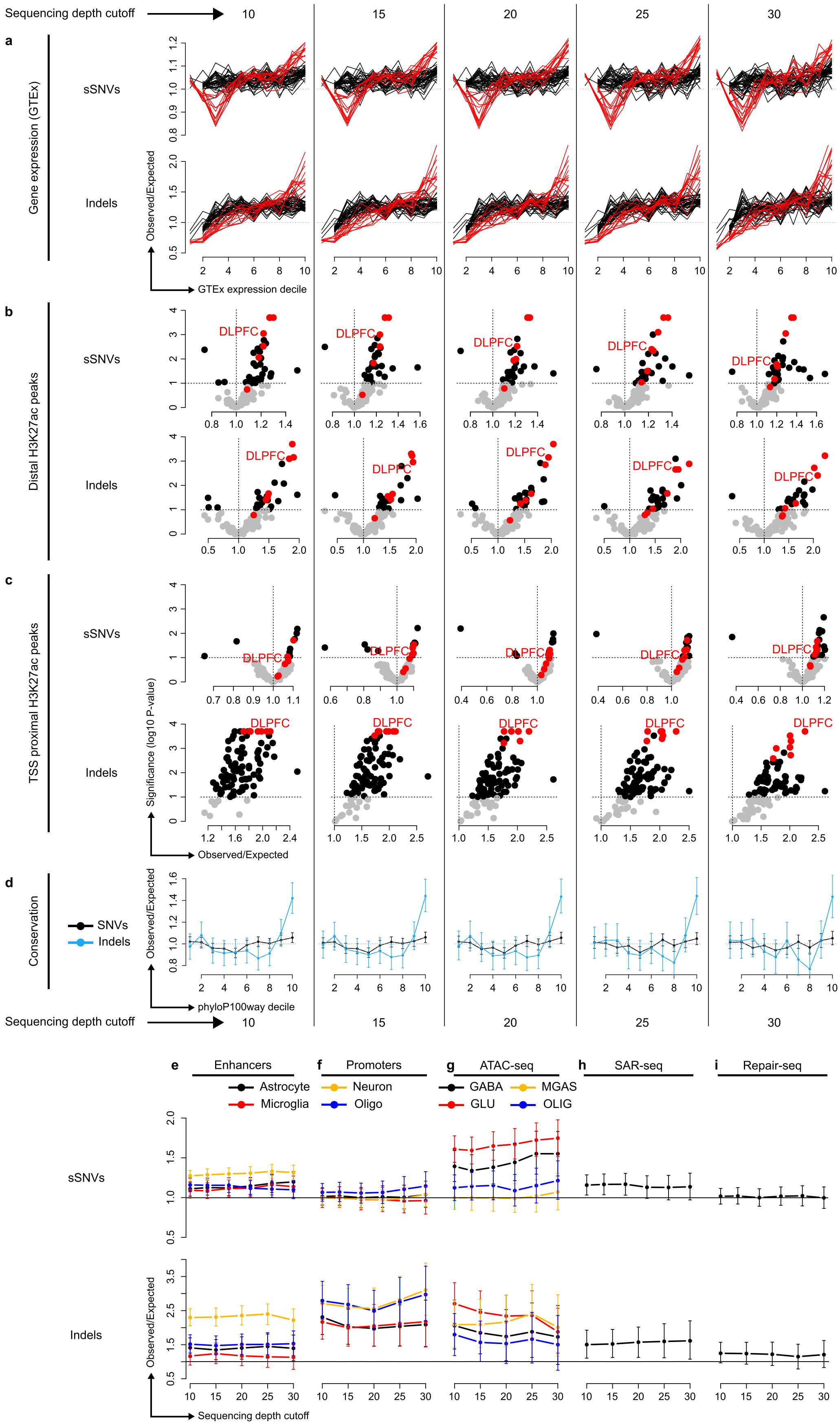


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | PTA sensitivity over genomic regions for SNVs and indels. a. Absolute sensitivity for spatial measurements that divide the genome into roughly equally sized deciles (median GTEx expression for a single tissue type, brain BA9 prefrontal cortex, and phyloP 100way conservation).
b-c. Relative sensitivities: sensitivity inside of the tested region divided by sensitivity of the complemented region. Enhancers and promoters from Nott et al. 2019, ATAC-seq from Hauberg et al. 2020, DNA repair hotspots from Wu et al. 2021 and Reid et al. 2021, H3K27ac peaks from Roadmap Epigenomics. Each point represents one PTA neuron; crosses represent the 7 PTA neurons sequenced to 60x, circles represent 30x depth samples. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median.



Extended Data Fig. 8 | ChromHMM states and neuronal mutations. Enrichment analysis of ChromHMM states from 127 tissues from the Roadmap Epigenomics Project. Active regions include 1_Tss, 4_Tx, 5_TxWk, 6_EnhG and 7_Enh; inactive states include 9_Het and 14_ReprPCWk. Red points, brain tissue regardless of significance level; black points, non-brain tissue; grey points, enrichment not significant at the P < 0.1 level. No correction for multiple hypothesis testing was applied.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Patterns of mutation enrichment persist at increasing sequencing depth thresholds. Analyses presented in Fig. 5 rerun using mutations supported by at least 10, 15, 20, 25 and 30 reads; permutations used for enrichment analysis are also restricted to the subset of the genome with the corresponding sequencing depth. GABA, GABAergic neurons; GLU, glutamatergic neurons; OLIG, oligodendrocytes; MGAS, microglia and astrocytes. Error bars: 95% bootstrapping confidence intervals. For panels **a**-**d**, each plot presents an analysis at one depth cutoff; for panels **e**-**i**, each plot contains the full range of depth cutoffs, as indicated on the x-axis. Error bars in **d**-**i** represent bootstrap 95% C.I.s using *n*=10,000 bootstrap samples; centre points are the observed mutation count divided by the mean mutation count of the bootstrap samples.

nature portfolio

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

Last updated by author(s): May 9, 2022

Reporting Summary

Nature Portfolio 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 Portfolio 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	Data was downloaded by Aspera ascp (version 2.7.0) and extracted with the SRA toolkit (version 2.10.7).
Data analysis	Data were analyzed using an environment with >300 packages, which were managed by Miniconda v3. The exact conda environment with all package versions is provided at https://github.com/parklab/SCAN2_PTA_paper_2022. SCAN2 v0.9 (both pipeline and the R package) was used; both of these are provided at the above Github repository. Additional packages not installed via conda are: MatLab v2019a, SigProfiler 2.5.1.7 and SigProfilerMatrixGenerator 1.1.9.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All MDA-amplified single neurons and matched bulks listed in Supplementary Table 2 were downloaded from dbGaP, identifier phs001485.v1.p1. Only prefrontal cortex neurons were used. Raw sequencing data for PTA-amplified single neurons has been uploaded to dbGaP, identifier phs001485.v3.p1, and PTA mouse ESCs have been uploaded to SRA, accession PRJNA832209.

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	No sample size calculations were performed. We compared yearly mutation accumulation rates with those estimated in Lodato et al, Aging and neurodegeneration are associated with increased mutations in single human neurons, Science 359, 555-559 (2018), in which 3-4 neurons per individual were sufficient to approximate the number of mutations accumulated per year.
Data exclusions	One MDA-amplified neuron, 5087pfc-Rp3C5, generated in Lodato et al. 2018, was excluded for having extremely high SNV burden. One PTA-amplified neuron generated in this study, 4638-Neuron-4, was excluded due to a very low mutation burden (both SNV and indel).
Replication	The nature of this study is to include all samples and cells that were examined, such that distinct cells from the same individual may considered biological replicates. Comparison to other datasets generated from different previous studies showed strong concordance (Lodato et al. 2018 Science, Xing et al. 2021 PNAS, Abascal et al. 2021 Nature)
Randomization	Not relevant: there is no case/control designation in this study.
Blinding	Not relevant: there is no case/control designation in this 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

Antibodies

Antibodies used	anti-NeuN (Millipore, MAB377X, clone A60, AlexaFluor-488 conjugated , 1:1250)
Validation	anti-NeuN: reactivity validated by the company for human. Validated by the company for FACS.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Hochepied et al. (2004), Stem Cells, 22(4): 441-447
Authentication	The cells were whole genome sequenced and shown to be of C57BL/6J x SPRET/Ei F1 background
Mycoplasma contamination	The cells were tested negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Human research participants

Policy information about studies involving human research participants

Population characteristics

We selected 17 individuals with no history of neurologic disease spanning a range of ages (0.4 - 104, mean 47.0), with 8 males and 9 females.

Recruitment

Tissue was obtained from participants in a brain donation program at the UMB NIH Neurobiobank.

Ethics oversight

Tissue collection and distribution for research and publication was conducted according to protocols approved by the University of Maryland Institutional Review Board (for UMBTB: 00042077), and after provision of written authorization and informed consent. Research on these de-identified specimens and data was performed at Boston Children's Hospital with approval from the Committee on Clinical Investigation (S07-02-0087 with waiver of authorization, exempt category 4).

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