This is the author's accepted manuscript of the article published in Nature Genetics. The final authenticated version is available online at: https://doi.org/10.1038/s41588-022-01180-2

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

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

[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; Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA; and Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[4] Division of Neuropathology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
[5] Center for Inflammation Research, VIB, Ghent, Belgium.
[6] Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium.
[7] Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea
[8] Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA, USA.
[9] Hubrecht Institute–KNAW, University Medical Center Utrecht, Uppsalalaan 8, 3584CT Utrecht, Netherlands[10] Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA.
[11] Chan Zuckerberg Biohub, San Francisco, CA, USA.
[12] BioSkryb, Durham, NC, USA.
[13] Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA.

[*] These authors contributed equally to this work.
[§] These authors jointly supervised this work.

Peter J. Park peter_park@hms.harvard.edu
Christopher A. Walsh Christopher.Walsh@childrens.harvard.edu

Abstract

Accurate somatic mutation detection from single-cell DNA sequencing (scDNA-seq) 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 non-clonal somatic mutation in single neurons with age, but revises the estimated rate of this accumulation to be 16 SNVs per year. We also identify artifacts in other amplification methods. Most importantly, we show that somatic indels increase by at least 3 indels 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 significant effect on genome integrity in human neurons.

Introduction

Although somatic mutation has been studied extensively in cancer, investigation into the abundance, patterns, and effects of somatic mosaicism in non-neoplastic 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 homogenous 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 (i.e., 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 post-mitotic cells such as neurons. A major challenge, however, is the difficulty of amplifying the genome of a single cell accurately and evenly prior to sequencing. For example, multiple displacement amplification (MDA)[12], a popular amplification method for detecting point mutations, produces non-uniformity 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 non-clonal and thus cannot be confirmed in other cells. We previously used LiRA[15], a read-level phasing strategy, 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 single nucleotide polymorphisms (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 called 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 PTA protocols from the prefrontal cortices of the same individuals shows that PTA substantially improves upon 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 order of magnitude more false positives (FPs) than there are mutations in some non-neoplastic 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 novel mutation signature[17] approach to increase sSNV detection sensitivity. Furthermore, SCAN2 enables analysis of somatic indels from single-cell DNA sequencing data for the first time. 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 non-clonal, private mutations such as those that occur in post-mitotic 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 prefrontal cortex (PFC) of 12 neurotypical individuals and sequenced to 30-60X, including 15 neurons from 5 neurotypical individuals from another study[20] (**Figure 1a, Supplementary Table 1**). 75 single neurons from 11 of the 17 individuals 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 to MDA-amplified cells, including substantial reduction in coverage variability and allelic dropout across the genome (**Figure 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 Figure 1**). Surprisingly, large-scale somatic copy number mutations (>5 Mb, Methods) were detected in only two of the 52 PTA neurons (**Supplementary Note** and **Supplementary Figure 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 (on the order of 10,000 per MDA amplification[23]) and indels, typically by spontaneous DNA damage or polymerase errors. The majority of 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 the majority of late artifacts. Early artifacts, especially those that occur prior to amplification (e.g., during cell lysis), can be more difficult to identify since 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: because both true mutations and SSD artifacts should have near-100% VAF, a systematic excess of near-100% VAF putative mutations in MDA compared to PTA neurons from the same individual would imply presence of SSD artifacts. Using a simple genotyping approach (Methods), we found a median excess of 15 somatic SNVs and 3.7 somatic indels in MDA X chromosomes (**Figure 2a-b, Supplementary Figure 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 arifacts, 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 vs. 26 sSNVs per neuron) compared to those from PTA-amplified neurons despite similar detection sensitivity, enriched for C>T mutations (85% vs. 59%; **Figure 2c**) and resembled two signatures previously reported to be associated with technical artifacts (Signature B[6] and Signature scF[24]; **Figure 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 mutations acquisition during developmental mitoses rather than an artifactual origin. Nevertheless, raw mutation counts indicate a much lower burden of SNV artifacts compared to MDA.

SCAN2: detecting somatic SNVs and indels in PTA single cells

SCAN2 builds upon 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 novel 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 (**Figure 3a**, Methods). First, the signature of true mutations is learned by “VAF-based” calling (which determines if 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 Figure 3**) and candidate mutations rejected in the first pass may be rescued if they are unlikely to have originated from the artifact signature (**Figure 3b**; see **Supplementary Figure 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 since it can strongly influence FDR. For high mutation burdens (e.g., germline variant detection), a genotyper’s FDR may appear low since 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 (e.g., 0.1-1.0 sSNVs/Mb[5,6,9,10], **Supplementary Note** and **Supplementary Figure 5**).

On simulated sSNVs, SCAN2 outperformed SCAN-SNV by increasing sensitivity by ~82% (46% vs. 25%) while maintaining similar FDR (8.6% vs. 9.5%) (**Extended Data Figure 2a,b**). SCAN2 also outperformed two other single-cell SNV genotypers (Monovar[18] and SCcaller[23]) by several-fold reduction of FDR (**Extended Data Figure 2c,d**). For SCAN2’s signature-based rescue, near-maximal performance was achieved when 500-1000 mutations were available for learning the mutation signature of true sSNVs (**Extended Data Figure 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 Figure 2g-i**).

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 validate by this approach, resulting in an overestimated FDR. We therefore used crossbred mouse embryonic stem cell (mESC) lines, which have greatly increased SNP density (~10-fold 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% vs. 26%) with FDR between 9%-32% depending on how false positives were defined (Methods, **Extended Data Figure 3a,b**). SCAN2 recovered 28% more sSNVs than SCAN-SNV on the AAI-treated clone (52% vs. 41%) and clearly recovered the aristolochic acid signature (**Extended Data Figure 3c**). On AAI-treated cells, both SCAN2 and SCAN-SNV achieved FDR $\approx$ 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 (**Figure 3c**). While it is rare for a particular sSNV artifact to occur twice in the same amplification, processes that generate artifactual indels (e.g., 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 (e.g., 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 which 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: naïve application of SCAN-SNV to indels yielded 19.9% sensitivity but 61%-85% of calls were false positives; 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 mean FDR <2% **Extended Data Figure 4a-c**). In contrast to sSNVs, indel properties such as their length often affect detection sensitivity; indeed, we found reduced sensitivity for indels in homopolymers and tandem repeats of >4 units (**Extended Data Figure 4d,e**).

The effects of various SCAN2 filtering steps on sSNV and indel calling are provided in **Supplementary Figure 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 Figure 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 (**Figure 4a**). LiRA, which predicts genome-wide mutations 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 Figure 5**). *De novo* signature analysis of VAF-based sSNVs from PTA-amplified neurons confirmed Signature A, an aging-associated signature we previously recovered from MDA-amplified neurons[6] (**Supplementary Figure 8**). Notably, no signature resembling Signature B was extracted from *de novo* analysis of our PTA sSNVs. Importantly, our filters, which require sSNVs 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 this 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/year to 22 sSNVs/year and removal of a strong elderly outlier (subject 5219) further decreased the rate to 19 sSNVs/year, more closely matching that of PTA neurons (**Supplementary Note** and **Supplementary Figure 9**). Taken together, these observations provide compelling evidence that sSNVs accumulate in human neurons at a rate closer to 16 sSNVs/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 ~3 somatic indels per neuron per year (Methods, **Figure 4b**), which is similar to rates observed in several mitotically active cell types[8-10,27]. However, our rate likely represents a lower bound on indel accumulation owing to lower sensitivity for indels of varying length and repeat content. Deletions accumulated 3.3-fold faster than insertions (**Figure 4c**) and indel sizes ranged from -29 bp to +17 bp (**Figure 4d**). As was the case for sSNVs, MDA yielded a higher accumulation rate estimate of 6.0 somatic indels/year and we again attribute this to MDA artifacts (**Supplementary Figure 10a**). 7 out of 75 MDA neurons contained an exceptionally high number of indel calls characterized by single base insertions in homopolymers of length 3 or greater (**Supplementary Figure 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 (**Figure 4e**) that was broadly similar to spectra from dividing cells[9,10,27] but with a greater burden of deletions (**Extended Data Figure 6a-e**). 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 (**Figure 4f**, **Extended Data Figure 6f**), 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 post-mitotic neurons. However, our analysis of indel sensitivity on simulated data indicated that lack of ID1 and ID2 could also be explained by low sensitivity that uniquely impacts these signatures (**Extended Data Figure 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 (**Figure 4g**, **Extended Data Figure 6g,h**); 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 vs. 14,748) and 76% more somatic indels (2,714 vs. 1,541) from PTA neurons compared to VAF-based calls. Only a handful of neurons showed evidence of deviation from the batch-wide sSNV and indel signatures ($P < 0.05$ for 3/52 and 2/52 neurons for SNV and indel signatures, respectively; statistical test described in **Supplementary Note** and **Supplementary Figure 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, **Extended Data Figure 7**).

Spurred by reports of transcriptional strand bias in neuronal SNVs (particularly T>C mutations in A$\underline{T}$N trinucleotide contexts[2,6]), we first compared neuronal somatic mutation density to gene expression levels from 54 tissues in 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 (**Figure 5a,b**), with the most expressed decile containing a ~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 8:1 (**Figure 5c**). Indels were also strongly enriched in the 10% of the genome with the highest evolutionary conservation, with an overrepresentation of 42% (**Figure 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 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 (**Figure 5e**). Near active TSSs, only somatic indels showed evidence of enrichment and it was not tissue specific (**Figure 5f**). Chromatin states[32]—which offer alternative definitions of promoters and enhancers based on a combination of chromatin marks from ChIP-seq signals—showed similar patterns, with indel enrichment in active TSSes (ChromHMM annotation: 1_Tss) and non-genic enhancers (7_Enh; **Extended Data Figure 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 which 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), while 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 (**Figure 5g**) but reduced or marginal significance in enhancers active in non-neuronal 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 (**Figure 5h**).

Analysis of open chromatin regions (OCRs) derived from ATAC-seq of flow-sorted GABAergic 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 while indel enrichment was strong but less tissue specific (**Figure 5i**).

Finally, we measured enrichment of mutations in the DNA repair hotspots recently reported by Wu et al. and Reid et al. (refs. 30 and 29). Enhancer-associated hotspots[30] were enriched for somatic indels (51% increase; 95% CI [11%, 90%], $P = 0.02$) but no enrichment was found in promoter-associated hotspots[29] (**Figure 5j**). sSNVs were also enriched in enhancers but with marginal significance. Notably, all enrichments presented in **Figure 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 Figure 9**).

Discussion

Our analyses of PTA-generated single-neuron genome sequencing represent a major advance in single-cell DNA-sequencing 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 post-mitotic human neurons. Further, SCAN2 analysis of 52 PTA neurons provided mutation density profiles that, 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. While PTA itself is a significant 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 (e.g., mutational signatures) to enhance sensitivity for non-shared sSNVs and indels genome-wide. Indeed, compared to 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 (e.g., *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 (e.g., within promoter or enhancer regions), we recommend correcting for local differences in nucleotide content to account for signature-related biases in the SCAN2 calls, as done in this study using permutations.

The rates and signatures of SNV and indel mutations 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 >6-fold more sSNVs than the META-CS study (~20,000 vs. ~3,000) and ~4-fold more sSNVs and indels than the NanoSeq study (~20,000 sSNVs and ~2,700 indels vs. ~5,000 sSNVs and 600 indels for this study and NanoSeq, respectively). Furthermore, the majority of the human genome is accessible to PTA while 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 only found an association between indel burden—not sSNV burden—and transcription levels and found 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 editing at the single cell level.

Acknowledgements

We thank R.S. Hill, Ronald Mathieu and Lakshmi (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 (IDDRC) 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 for the advancement of science. This work was supported by the Bioinformatics and Integrative Genomics training grant (L.J.L., T32HG002295), K08 AG065502 (M.B.M.), T32 HL007627 (M.B.M.), the Brigham and Women’s Hospital Program for Interdisciplinary Neuroscience through a gift from Lawrence and Tiina Rand (M.B.M.), the donors of the Alzheimer’s Disease Research program of the BrightFocus Foundation A20201292F (M.B.M.), the Doris Duke Charitable Foundation Clinical Scientist Development Award 2021183 (M.B.M.), PRMRP Discovery Award W81XWH2010028 (Z.Z.), the Edward R. and Anne G. Lefler Center postdoctoral fellowship (Z.Z.), R00 AG054748 (M.A.L.), R01 AG070921 (C.A.W.), R01NS032457, U01MH106883 (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 (C.A.W.). C.A.W. is an Investigator of the Howard Hughes Medical Institute. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions. LJL conceived and implemented SCAN2. CAW, MBM, and ZZ conceived the application of PTA to single neurons. PJP and CAW conceived and supervised the overall project. LJL, MBM and ZZ wrote the manuscript; all authors reviewed and edited the manuscript. LJL and YZ performed computational analysis. LJL, MBM and ZZ analyzed and interpreted results. MBM, ZZ, JG, SB, SK and MAL collected tissue specimens, isolated single neuronal nuclei and performed PTA amplification and amplification quality control studies. CLB and LJL performed LiRA analysis and comparisons to SCAN2. TH and CL generated mESC. JIG conceived and performed the mESC kindred experiment. CLB, AG and JK collected and processed all sequencing data. DG and HJ made suggestions for signature analysis. CG and JW 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 this 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 this study. The remaining authors declare no competing interests.

Main Figure Legends

Figure 1. Improved large-scale amplification characteristics of PTA compared to MDA. **a.** Study design. Single neurons were collected from the prefrontal cortex (PFC) of brains of 17 individuals ranging in age from infantile to elderly. Single neurons were amplified by either PTA or MDA and then sequenced to high coverage. Created with BioRender.com. **b.** Representative copy number profiles for bulk (top), MDA-amplified (middle) and PTA-amplified (bottom) genomes. **c.** MAPD (median absolute pairwise difference) 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. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median. *n*=17 bulk samples, *n*=52 PTA neurons, *n*=75 MDA neurons. **d.** Allele balance for germline heterozygous SNPs measures 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 0.3-0.7) compared to only 39% of each MDA genome.

Figure 2. PTA identifies MDA-induced artifacts.

a-b. Sensitivity-adjusted somatic SNV (sSNV) (**a**) and indel (**b**) burdens per X chromosome for 5 male subjects with both MDA and PTA-amplified neurons. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median; *n*=16 PTA neurons and *n*=39 MDA neurons. **c.** Fraction of C>Ts among SCAN-SNV sSNV calls in infant neurons and two previously published signatures. **d.** Mutation spectra of SCAN-SNV sSNVs across 13 MDA infant neurons, 6 PTA infant neurons, the C>T rich Signature B reported in Lodato et al, 2018 and the MDA artifact Signature scF reported by Petljak et al, 2019. Light red bars denote C>Ts that occur at CpG sites. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median.

Figure 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 used in 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 in panel (b). 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*. **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.

Figure 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. **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 Figure 4f**) before fitting. **g.** Association of ID4, a signature of unknown aetiology, with neuron age; *P*-value: two-sided *t*-test for correlation=0. Trend lines in a-c and g: mixed effects linear regressions to account for multiple points being derived from the same individual.

Figure 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 to local gene expression levels measured by GTEx. Each line corresponds to one GTEx tissue type; tissues from primary brain specimens are always shown in red. **c.** The number of high impact (classified HIGH by SnpEff; includes several severely protein altering effects such as stop gains, stop losses and frameshifts) sSNVs and somatic indels detected by SCAN2's signature-based approach (dark grey) and extrapolation to autosomes (light grey). **d.** Mutation enrichment compared to 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. **g-j.** Mutation enrichment analysis of several datasets. Dorsolateral prefrontal cortex is shown since it most closely matches the neurons sequenced in this study. Cell-type specific enhancers (**g**) and promoters (**h**) from Nott et al. 2019; cell-type specific open chromatin regions (OCRs) measured by ATAC-seq from Hauberg et al. 2020 (**i**); DNA repair hotspots measured in induced human neurons (**j**) reported by Wu et al. 2021 (SAR-seq) and Reid et al. 2021 (Repair-seq). GABA, GABAergic neurons; GLU, glutamatergic neurons; OLIG, oligodendrocytes; MGAS, microglia and astrocytes. Error bars (**g-j**): 95% bootstrapping C.I. with $n=10^4$ bootstrap samplings; centre point: 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.

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., de Ligt, J., Jager, M. 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., Helgadottir, H. T. *et al*. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biology* **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., Adalsteinsson, V.A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K.L., Meyerson, M. & Love, J.C. 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., Natarajan S. *et al*. Accurate Genomic Variant Detection in Single Cells with Primary Template-Directed Amplification. *Proc. Natl. Acad. Sci. U.S.A.* **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 Meth* **13**, 505-507 (2016).
19. Singer, J., Kuipers, J., Jahn, K. and Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nat Commun* **9**, 5144 (2018).
20. Miller, M. B., Huang, A. Y., Kim, J., Zhou, Z., *et al*. Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714-722 (2022).

21. McConnell, M. J., Lindberg, M. R., Brennand, K. J., *et al*. Mosaic copy number variation in human neurons. *Science* **342**, 632-637 (2013).
22. Chronister, W. D., Burbulis, I. E., Wierman, M. B., *et al*. Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep* **26**, 825-835 (2019).
23. Dong, X., Zhang, L., Milholland, B., *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. *bioRxiv* doi: 10.1101/043448 (2016).
26. Lasken, R. S., Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology* **7**, doi: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., Platts, A., Wang LL, Coon, M. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* (Austin) **6**, 80-92 (2012).
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.* **118**, e2013106118 (2021).
37. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405-410 (2021).

Methods

Human tissue, case selection and ethical approval

Postmortem frozen human tissues were obtained from the NIH Neurobiobank at the University of Maryland Brain and Tissue Bank (UMBTB). 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). and processed according to an IRB-approved protocol at Boston Children's Hospital. Consent was obtained by the NIH Neurobiobank. Non-disease neurotypical individuals had no clinical history of neurologic disease and were selected to represent a range of ages from infancy to older adulthood.

Isolation of single neuronal nuclei for single-cell whole genome sequencing

Single neuronal nuclei were isolated using fluorescence-activated nuclear sorting (FANS) for NeuN, 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 (10mM Tris-HCl, 0.32M sucrose, 3mM Mg(Oac)2, 5mM CaCl2, 0.1mM EDTA, 1mM DTT, 0.1% Triton X-100, pH 8) on ice. Tissue lysates were carefully layered on top of a sucrose cushion buffer (1.8M sucrose 3mM Mg(Oac)2, 10mM Tris-HCl, 1mM DTT, pH 8) and ultra-centrifuged for 1 hour at 30,000 x g. Nuclear pellets were incubated and resuspended in ice-cold PBS supplemented with 3mM MgCl2, filtered (40 µm pore size), then stained with Alexa Fluor 488-conjugated anti-NeuN antibody (Millipore MAB377X). Large neuronal nuclei were then subjected to FANS, one nucleus per well into 96-well plates.

Single nucleus whole genome amplification by primary template-directed amplification (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, Durham, NC), using the manufacturer's protocol. Briefly, single nuclei were sorted into wells containing 3 µL Cell Buffer pre-chilled on ice, then alkaline lysed on ice with MS Mix, mixed at 1400rpm, 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 were added, then the amplification reaction was carried out for 10 hrs. at 30°C, followed by enzyme inactivation at 65°C for 3 min. Amplified DNA was then cleaned up using AMPure beads, and yield determined by the picogreen method (Quant-iT dsDNA Assay Kit, ThermoFisher). Samples were subjected to quality control by multiplex PCR for 4 random genomic loci as previously described[6], and by Bioanalyzer for fragment size distribution. Amplified genomes demonstrating positive amplification for all 4 loci were then prepared for Illumina sequencing. The majority of the PTA scWGS neuron experiments described here were performed specifically for this report, and they are supplemented with experiments from aged 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, 07141548001). Ligated DNA was cleaned up using AMPure beads and amplified through an on-bead PCR amplification. Amplified libraries were selected for 300-600 bp size using AMPure beads. Libraries were subjected to quality control using picogreen and Tapestation HS D1000 Screen Tape (Agilent PN 5067-5584) before sequencing. Single cell genome libraries were sequenced on the Illumina NovaSeq platform (150 bp x 2) at 30X except for subjects 1278 (HiSeq, 60X) and 1465 (NovaSeq, 60X). Illumina reads were aligned to the human reference with decoy sequence GRCh37d5 (hs37d5) using bwa mem.

Kindred mouse embryonic stem cell clones

Pluripotent mESCs on a C57BL/6J x SPRET/Ei F1 background were grown on feeders and maintained in N2B27 media supplemented with the glycogen synthase kinase-3 inhibitor, CHIR99021 (Axon Medchem, 1386, 3 µM), the MEK/ERK inhibitor PD0325901 (Axon Medchem, 1408, 0.4 µM), and mouse leukemia inhibitory factor (LIF) at 1000 U/ml referred to as to 2i + LIF media. mESCs were treated (or not) with aristolochic acid I 50 uM (AAI, Sigma A5512) for 48 hours, and subsequently disaggregated into single cells and plated at limiting dilution. Single cell clones were picked after one 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 30x on the Illumina NovaSeq platform (150 bp x 2) and aligned to GRCm38 using bwa mem.

Single-cell amplification quality metrics

Median absolute pairwise differences (MAPD) were computed by estimating copy number in bins CN_i of size 50 kb following ref. 39; subsequently, $\mathrm{MAPD} = \mathrm{median}(|\log_2 CN_i - \log_2 CN_{i+1}|)$. Copy number profiles in **Figure 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 3 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 hSNPs for each cell (located in path/to/SCAN2_output/ab_model/[single_cell]/hsnps.tab) were mapped to 1 kb non-overlapping tiles across autosomes from GRCh37d5. The allele balance for tile *i* containing hSNPs $\{j\}$ is $A_i = \sum_i H_{j,1}/(H_{j,1} + H_{j,2})$, where $H_{j,k}$ is the number of reads supporting haplotype *k*. The heatmap in **Extended Data Figure 1e** was produced by `pheatmap` with default parameters on the correlation matrix of *A* vectors.

Comparison of MDA and PTA somatic mutation calls

Both MDA- and PTA-amplified neurons were available for 5 male subjects. For X chromosome analysis, GATK HaplotypeCaller (v3.8.1) was run in joint mode across all samples (bulk, PTA and MDA) for each individual using dbSNP 147_b37_common_all_20160601 and parameters `--dontUseSoftClippedBases -rf BadCigar -mmq60`. GVCF joint calling was not used because information can be lost compared to 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 alternate 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 2 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 alternate reads in bulk and absence from dbSNP). The final estimated number of mutations was calculated by (#corrected calls) = (#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 or more mutation supporting reads across the 63 SDs, single-cell read depth >= 10; depth >=10, 0 mutation supporting reads and a 0/0 GATK genotype string in the matched synthetic bulk. Loci present in dbSNP v147_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 synthetic diploids (SDs) and the matched synthetic bulk with the same parameters as in section *Somatic mutation calling on male X chromosomes*. 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 was then run using 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

Synthetic diploid (SDs) 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. 9 SDs with 30x mean depth were generated by making all pairings of the 3 PTA cells from subjects 1278 and 5817 and by downsampling the reads in each BAM to ~15x. The youngest subjects (0.4 and 0.6 years old) were chosen to minimize the number of endogenous somatic mutations. Endogenous mutations were identified by applying GATK HaplotypeCaller v3.8 jointly to the 9 SDs, 6 original PTA BAMs and 2 matched bulks using the same parameters as in *Somatic mutation calling on male X chromosomes*. 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 >= 90% and <2 reference reads; depth >= 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 1000 SNV and indel spike-ins, for a total of 63 SDs. Random spike-in positions were uniformly sampled from chrX excluding assembly gaps (https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gap.txt.gz), 5 bp windows centered on each non-reference site reported by GATK in subject 1278 or 5817 and 5 bp windows centered on all sites in dbSNP v147 common. Somatic SNV spikeins 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 only created 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 >1000 candidates were obtained for each ID83 class. Somatic indel spikeins were further required to be >150 bp from the nearest indel spikein candidate to prevent crowding in repetitive tracts. SNV and indel spikeins 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 a false positive. 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 version 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 >= 10 with at least 3 mutation supporting reads; bulk depth >= 6 and <= 1 mutation supporting read; and single cell VAF ≥ 10% for sSNVs with >100 depth or VAF ≥ 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 version 1.1 was run following the authors recommendations. BAMs were converted to pileups using samtools version 1.3.1 with the option -C50 and hSNPs were defined using dbSNP version 147 common. Single cell somatic SNVs were called by applying SCcaller's `-a varcall`, `-a cutoff` and reasoning v1.0 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 $\alpha = 0.05$, while the stringent calling parameter is $\alpha = 0.01$.

SNV calling with LiRA

LiRA version `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 to use a single genome size for burden estimation. Current LiRA versions use 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 S5 of ref. 6 did not require similar correction.

Kindred mESC analysis

LiRA version `3bc0ae1` was used with the global option `reference_identifier GRCm38` and the `--force` flag to `lira varcall` following the authors' instructions. 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 custom dbSNP database of SPRET_EiJ sites from mgp.v5.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 non-simulated truth set. It was therefore necessary to reduce the high SNP density of cross-bred mice (~33 million SNPs/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 randomly 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 >=20% in the kindred clone bulk; >=5 reads in another kindred cell; or >=1 read in >=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% <= VAF < 100% and depth >= 10 in the kindred clone being analyzed; and VAF = 0 in the other kindred clone. A total of 130 clonal SNVs were 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 4 cells from each clone.

SCAN2 analysis of single human neurons

SCAN2 version 0.9 was run separately for each of the 17 subjects; for each subject, all MDA, PTA and bulk samples were provided to SCAN2. Non-default 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 version 147_b37_common_all_20160601. All following scan2 commands used SCAN2 v1.0. 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 here), and PTA neuron 4638-Neuron-4, due to a very low mutation burden.

Per-cell total mutation 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[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 version 3, signatures ID1-17. For the discovery of active signatures in **Figure 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 synthetic diploid X chromosomes (**Extended Data Figure 4f**).

Functional impact of point mutations

The severity of somatic SNV and indel mutations reported in **Figure 5c** were derived from SnpEff[29] version 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 (i.e., exact recurrence of a mutation) were either removed or downsampled to 1 call. For duplicate calls occurring in >1 subject, all instances were removed; for duplicate mutations in >1 neuron from the same subject (1.1% of sSNVs, 0% of indels), 1 occurrence was arbitrarily retained. An additional 57 sSNV calls were removed due to duplicates observations (not SCAN2 calls) in >1 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 v1.0) 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 permuted 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). Permuted 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 permuted 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. Confidence intervals for enrichment estimates are computed by bootstrapping the observed mutation set and computing enrichment as described above 10,000 times. To analyze enrichments with higher sequencing depth cutoffs *D* (**Extended Data Figure 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 >=*D*.

Genomic covariates for enrichment analysis

GTEx expression values for 54 tissues was 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 non-overlapping 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 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 intergenic. Finally, windows were ranked into deciles by their genic coverage-weighted TPM values and windows within each decile were merged to create 10 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 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 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` v2 program. Bins were then ranked into deciles by average phyloP score and windows within each decile were merged to create 10 regions. Cell-type specific enhancer and promoter regions[33] were extracted from Supplementary Table 5 tabs Astrocyte enhancers, Astrocyte promoters, etc.. Enhancer or promoter regions were merged within each cell type to produce 2 regions per cell type. Open chromatin regions for GABA, GLU, OLIG and MGAS from dorsolateral prefrontal cortex (DLPFC)[34] were downloaded from https://bendlj01.u.hpc.mssm.edu/ggoma/. SAR-seq DNA repair hotspots[30] were downloaded from GEO (GSE167257, GSE167257_SARseq_iNeuron_OverlapRep123.peaks.bed.gz); Repair-seq peaks[29] were obtained from Supplementary Table S1 of ref. 28.

Data availability

All MDA-amplified single neurons and matched bulks listed in **Supplementary Table 2** were downloaded from dbGaP, identifier phs001485.v1.p1. Only neurons from the pre-frontal corteces from individuals for which additional PTA data were generated were used. Raw sequencing read data for PTA-amplified human neurons can be downloaded from dbGaP, identifier phs001485.v3.p1. PTA-amplified mESC kindred cells and bulks can be downloaded from the NCBI Sequence Read Archive, identifier PRJNA832209.

Code availability

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

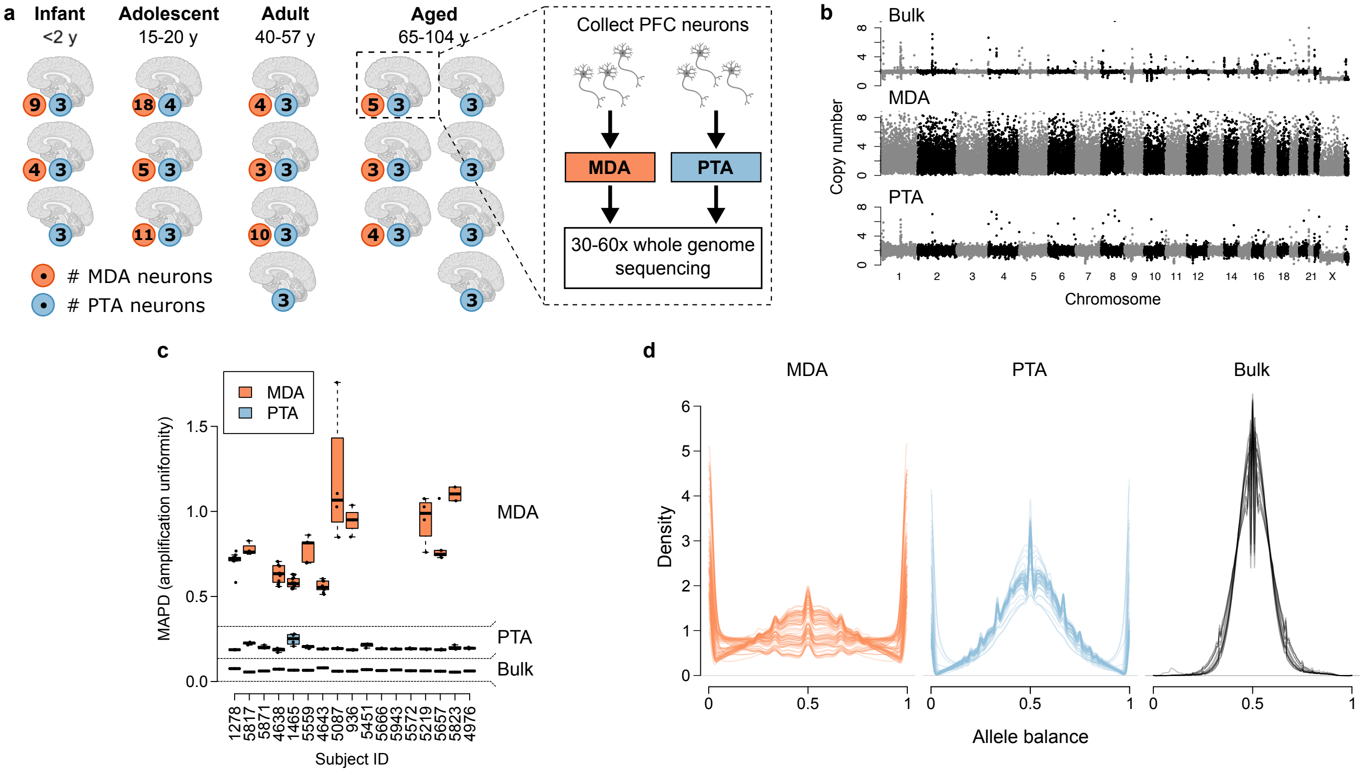
Statistics and Reproducibility

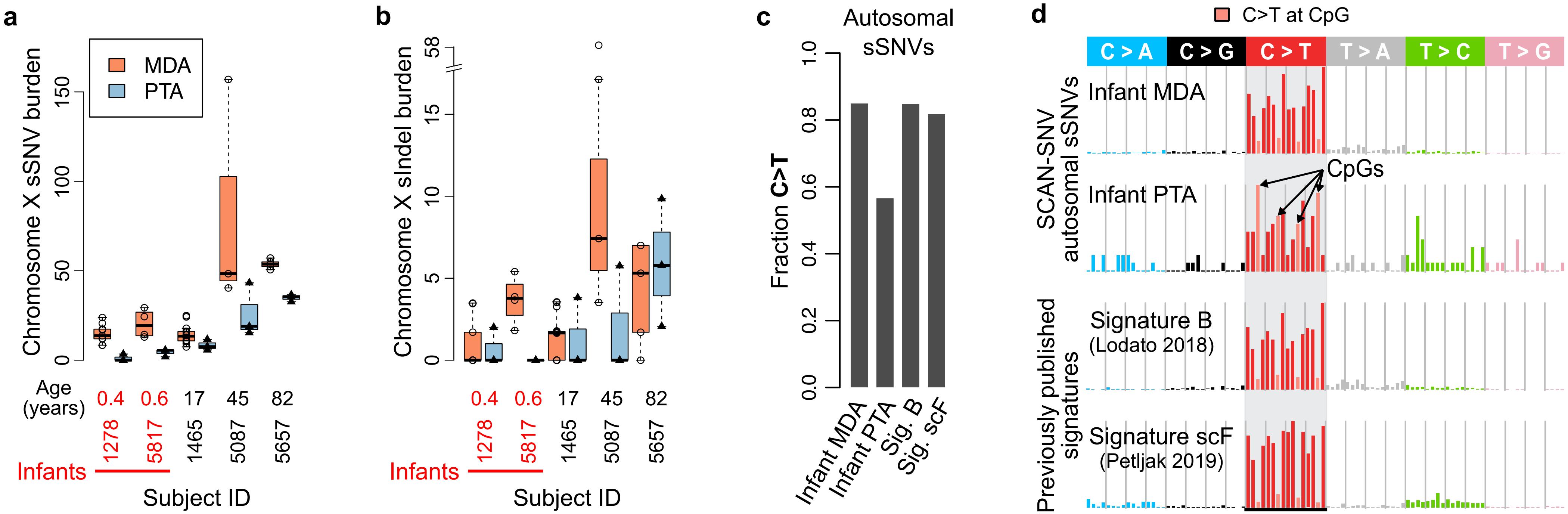
No statistical method was used to predetermine sample size. All 4 PTA neurons from brain 1465 were excluded from copy number analyses, but no other PTA neurons were excluded from any analysis. One MDA neuron from a previous study[6] (5087pfc-Rp3C5) was excluded from most analyses due to high mutation burden; one PTA neuron from this study (4638-Neuron-4) was excluded due to a very low mutation burden. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

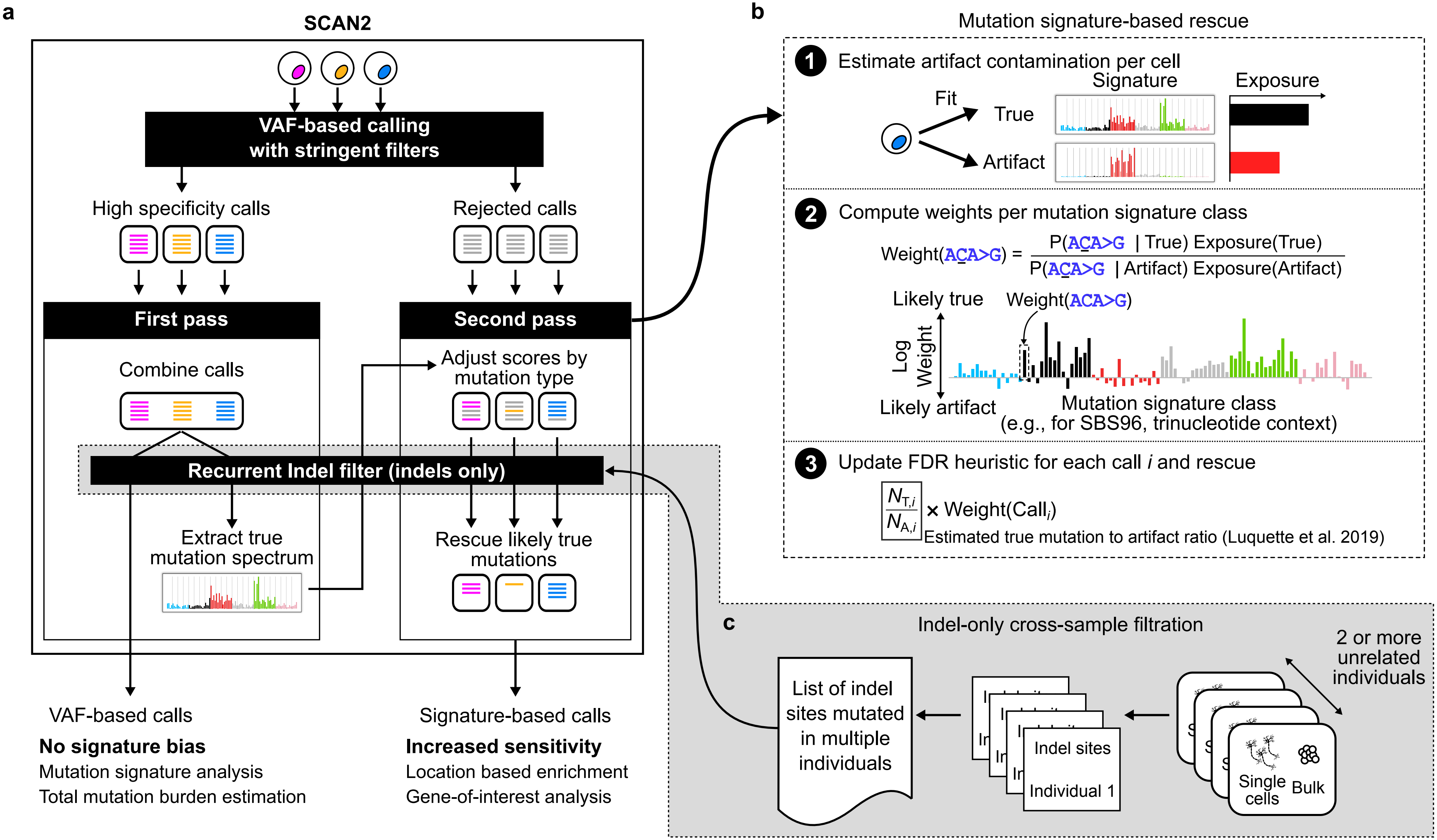
References

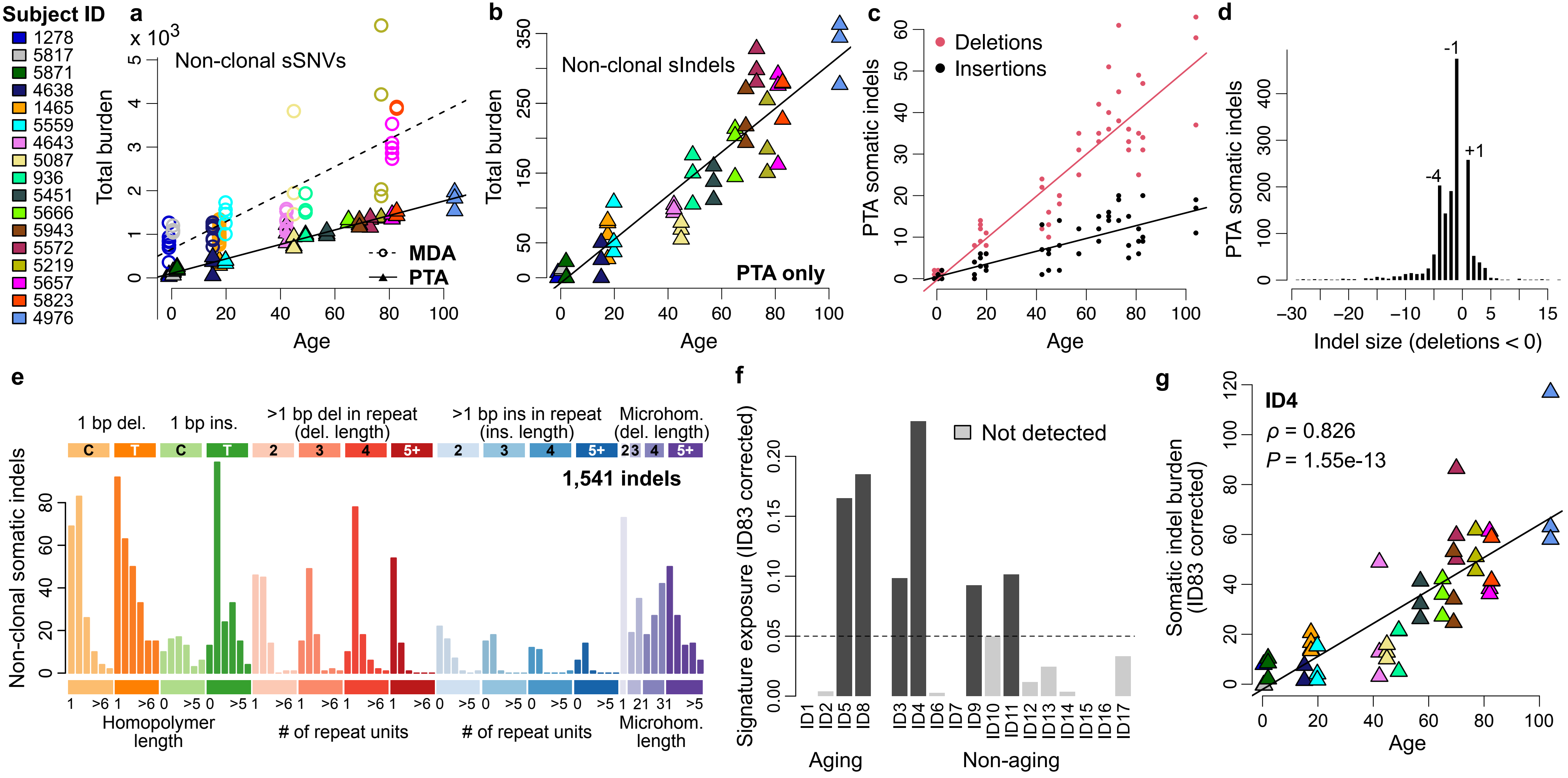
38. Evrony, G.D., Cai, X., Lee, E., *et al*. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).
39. Baslan, T., Kendall, J., Rodgers, L., et al. Genome-wide copy number analysis of single cells. Nat Protoc 7, 1024-1041 (2012).
40. Garvin, T., Aboukhalil, R., Kendall, J., et al. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods 12, 1058-1060 (2015).
41. Bergstrom, E. N., Huang, M. N., Mahto, U. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
42. Bates, D., Mächler, M., Bolker, B. and Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
43. Alexandrov, L. (2020). SigProfiler (https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler), MATLAB Central File Exchange. Retrieved January 1, 2020.
44. Luquette, L. (2022). SCAN2_PTA_paper_2022. Zenodo, doi:10.5281/zenodo.6532827.

999

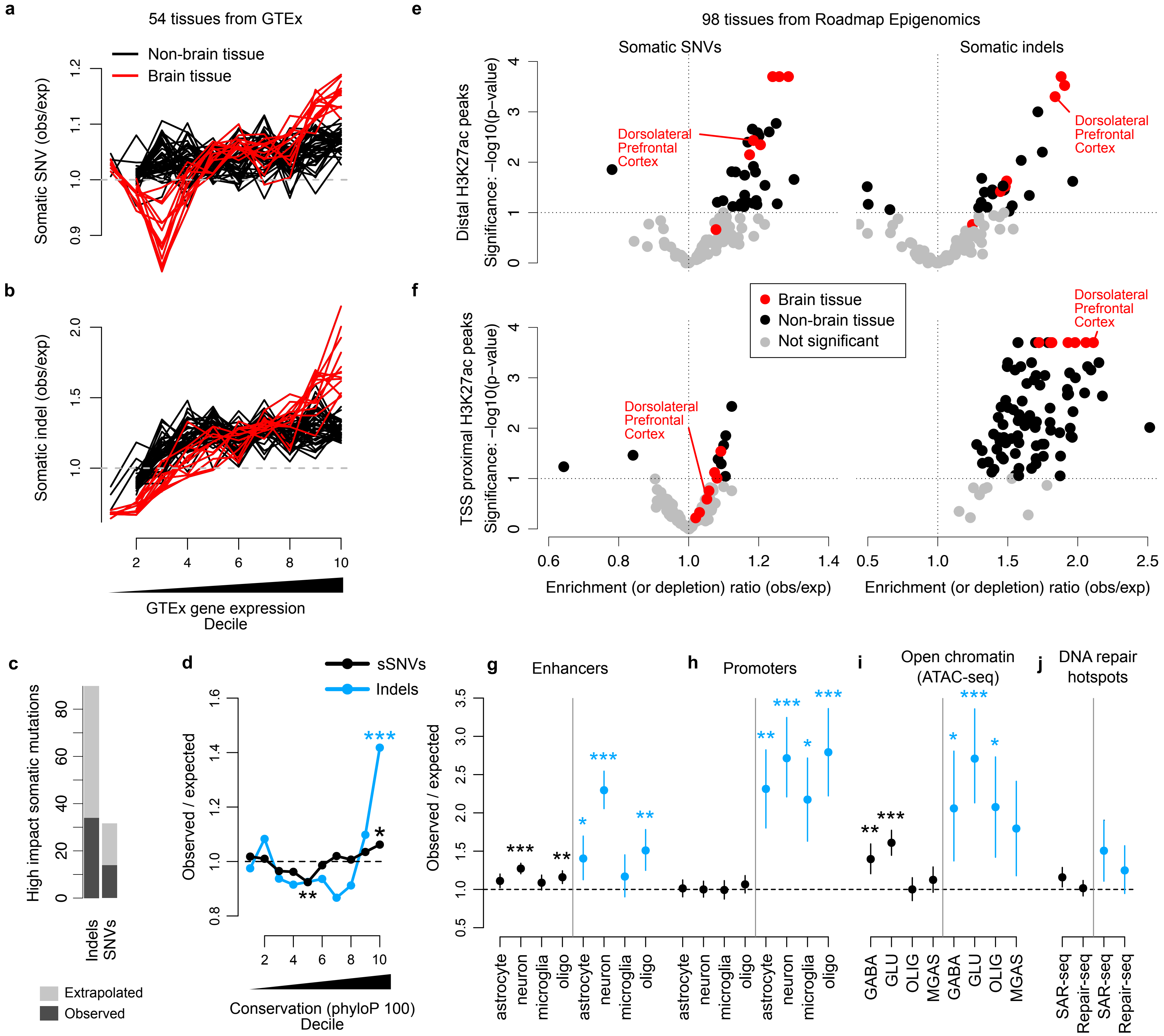

a
Infant
<2 y
Adolescent
15-20 y
Adult
40-57 y
Aged
65-104 y
9
3
4
3
3
18
4
5
3
11
3
4
3
3
3
10
3
3
5
3
3
3
3
4
3
3
3
3
3
# MDA neurons
# PTA neurons
Collect PFC neurons
MDA
PTA
30-60x whole genome sequencing
b
Bulk
MDA
PTA
Copy number
0
4
8
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
21
X
Chromosome
c
MDA
PTA
MAPD (amplification uniformity)
0.0
0.5
1.0
1.5
MDA
PTA
Bulk
1278
5817
5871
4638
1465
5559
4643
5087
936
5451
5666
5943
5572
5219
5657
5823
4976
Subject ID
d
MDA
PTA
Bulk
Density
0
1
2
3
4
5
6
0
0.5
1
Allele balance

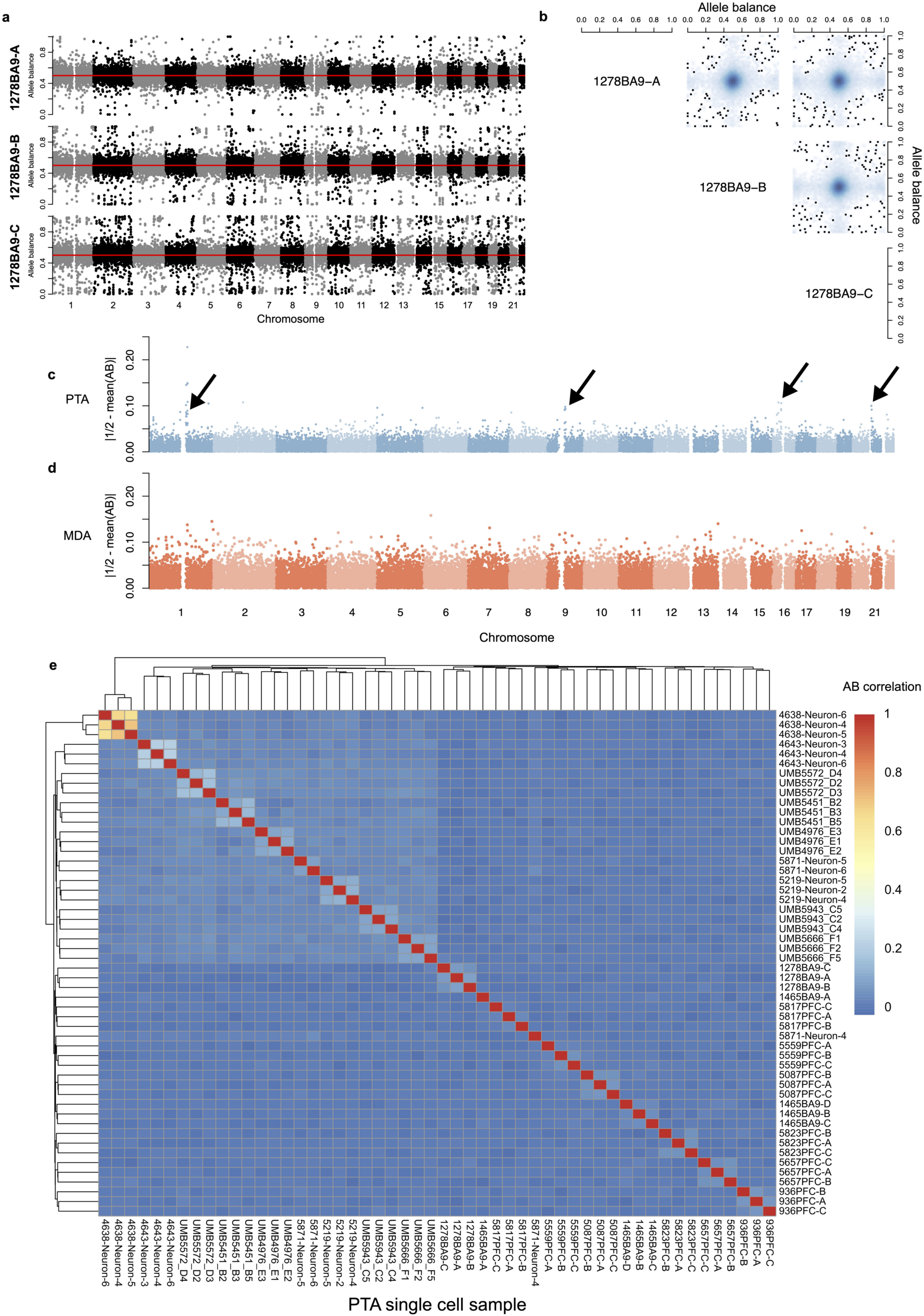

a
MDA
PTA
Chromosome X sSNV burden
0
50
100
150
Age (years)
0.4
0.6
17
45
82
1278
5817
1465
5087
5657
Infants
Subject ID
b
Chromosome X sIndel burden
0
5
10
15
58
0.4
0.6
17
45
82
1278
5817
1465
5087
5657
Infants
Subject ID
c
Autosomal sSNVs
Fraction C>T
0.0
0.2
0.4
0.6
0.8
1.0
Infant MDA
Infant PTA
Sig. B
Sig. scF
d
C>T at CpG
C > A
C > G
C > T
T > A
T > C
T > G
SCAN-SNV autosomal sSNVs
Infant MDA
Infant PTA
CpGs
Previously published signatures
Signature B (Lodato 2018)
Signature scF (Petljak 2019)

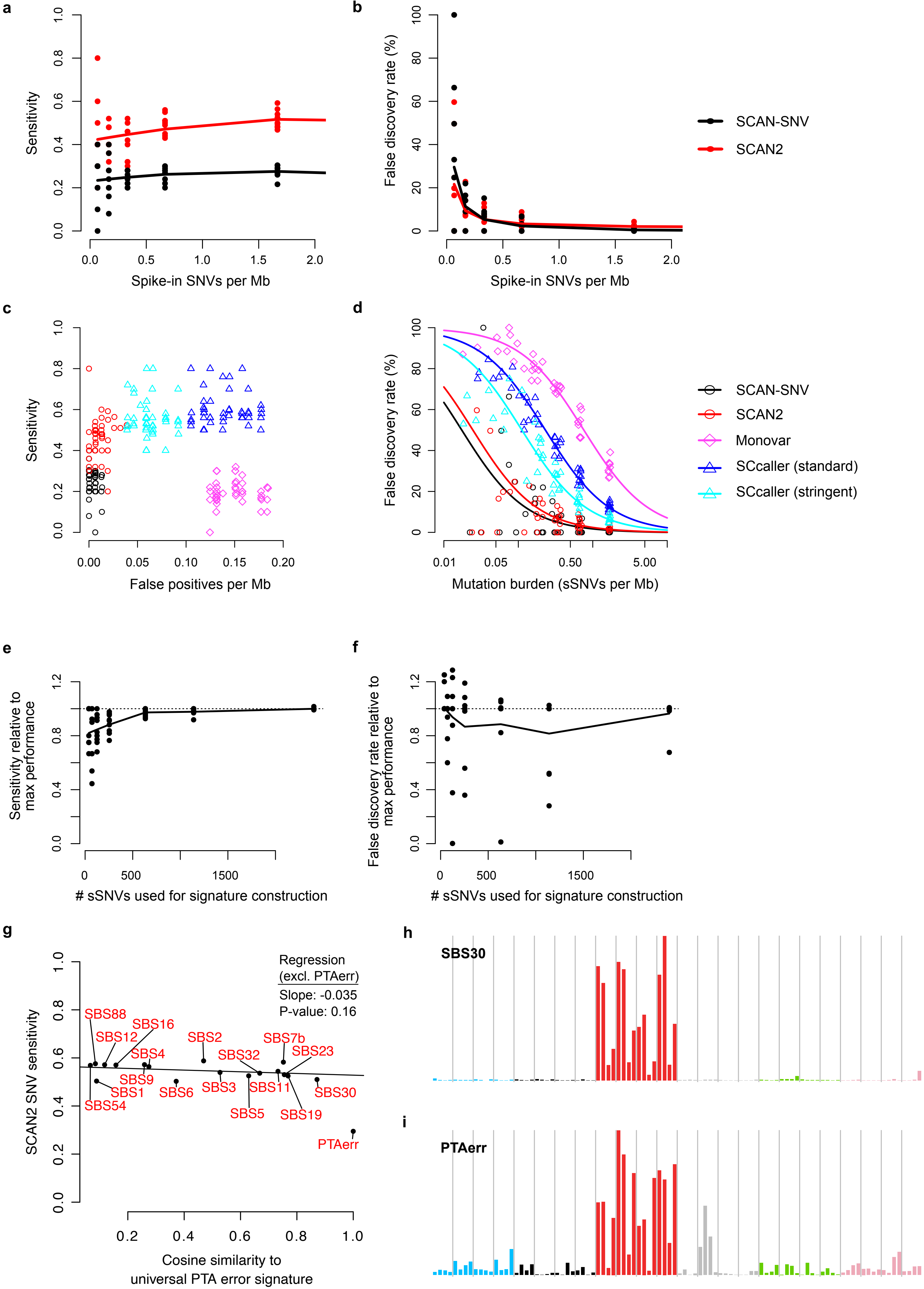

a
SCAN2
VAF-based calling
with stringent filters
High specificity calls
Rejected calls
First pass
Second pass
Combine calls
Adjust scores by
mutation type
Recurrent Indel filter (indels only)
Extract true
mutation spectrum
Rescue likely true
mutations
VAF-based calls
No signature bias
Mutation signature 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
(e.g., for SBS96, trinucleotide context)
3 Update FDR heuristic for each call i and rescue
$\frac{N_{T,i}}{N_{A,i}}$ × Weight($Call_i$)
Estimated true mutation to artifact ratio (Luquette et al. 2019)
c
Indel-only cross-sample filtration
List of indel
sites mutated
in multiple
individuals
Indel sites
Individual 1
2 or more
unrelated
individuals
Single
cells
Bulk

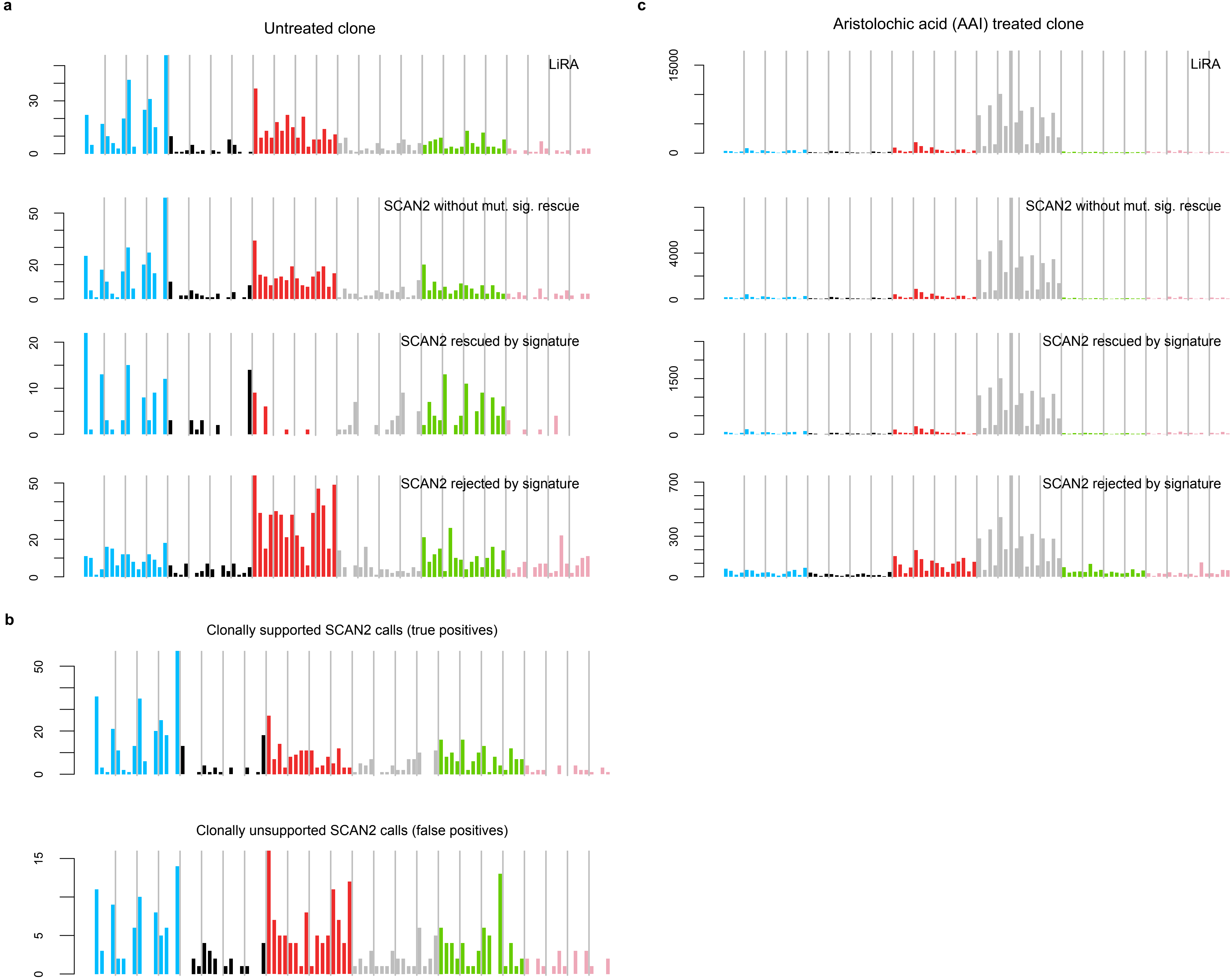


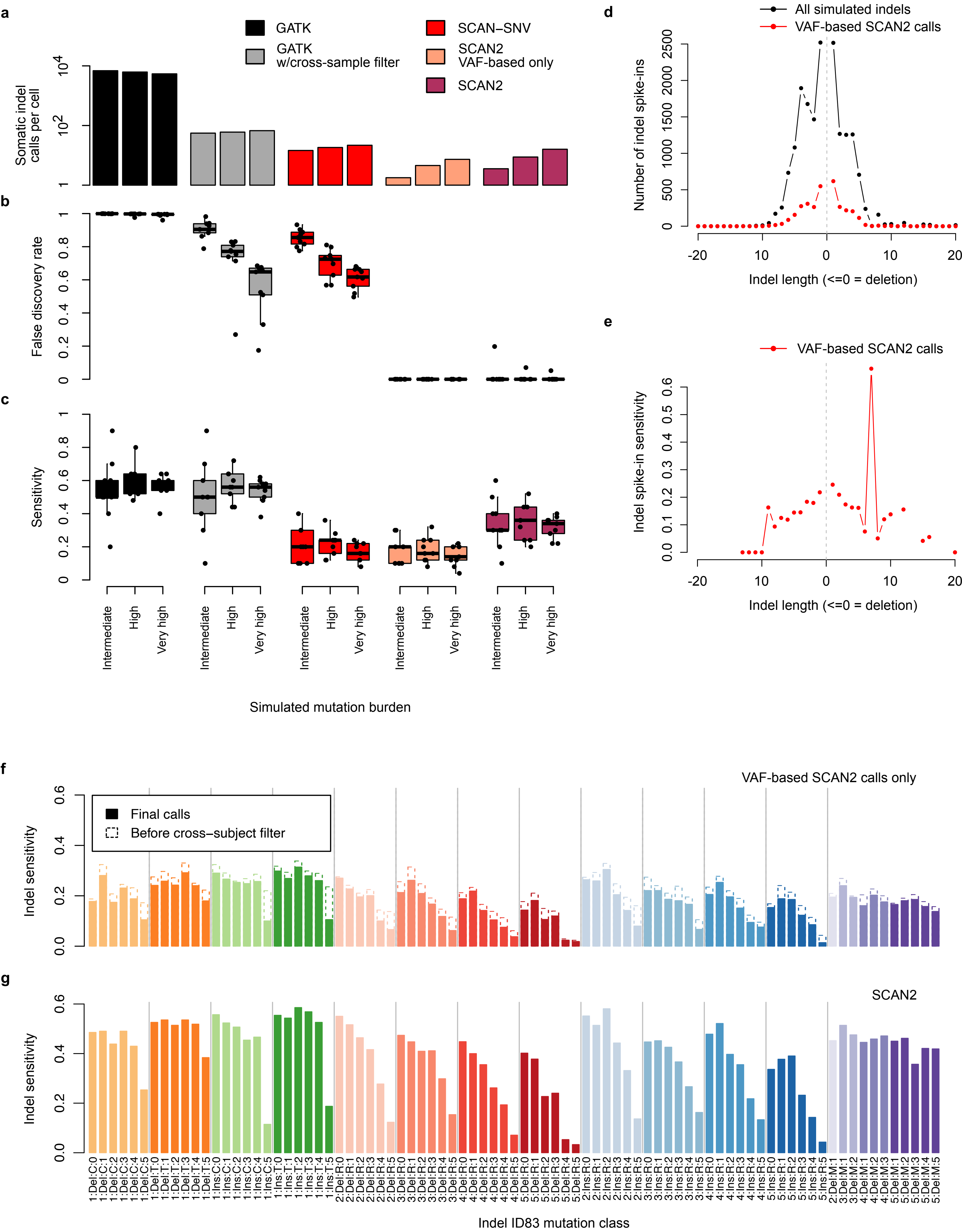
Subject ID
1278
5817
5871
4638
1465
5559
4643
5087
936
5451
5666
5943
5572
5219
5657
5823
4976
a
x 10^3
Non-clonal sSNVs
Total burden
0
1
2
3
4
5
0
20
40
60
80
100
Age
MDA
PTA
b
Non-clonal sIndels
Total burden
0
50
150
250
350
0
20
40
60
80
100
Age
PTA only
c
Deletions
Insertions
PTA somatic indels
0
10
20
30
40
50
60
0
20
40
60
80
100
Age
d
PTA somatic indels
0
100
200
300
400
-1
-4
+1
–30
–20
–10
0
5
15
Indel size (deletions < 0)
e
1 bp del.
1 bp ins.
>1 bp del in repeat
(del. length)
>1 bp ins in repeat
(ins. length)
Microhom.
(del. length)
C
T
C
T
2
3
4
5+
2
3
4
5+
2
3
4
5+
1,541 indels
Non-clonal somatic indels
0
20
40
60
80
1
>6
1
>6
0
>5
0
>5
1
>6
1
>6
1
>6
1
>6
0
>5
0
>5
0
>5
0
>5
1
21
31
>5
Homopolymer
length
# of repeat units
# of repeat units
Microhom.
length
f
Signature exposure (ID83 corrected)
0.00
0.05
0.10
0.15
0.20
Not detected
ID1
ID2
ID5
ID8
ID3
ID4
ID6
ID7
ID9
ID10
ID11
ID12
ID13
ID14
ID15
ID16
ID17
Aging
Non-aging
g
ID4
ρ = 0.826
P = 1.55e-13
Somatic indel burden
(ID83 corrected)
0
20
40
60
80
100
120
0
20
40
60
80
100
Age

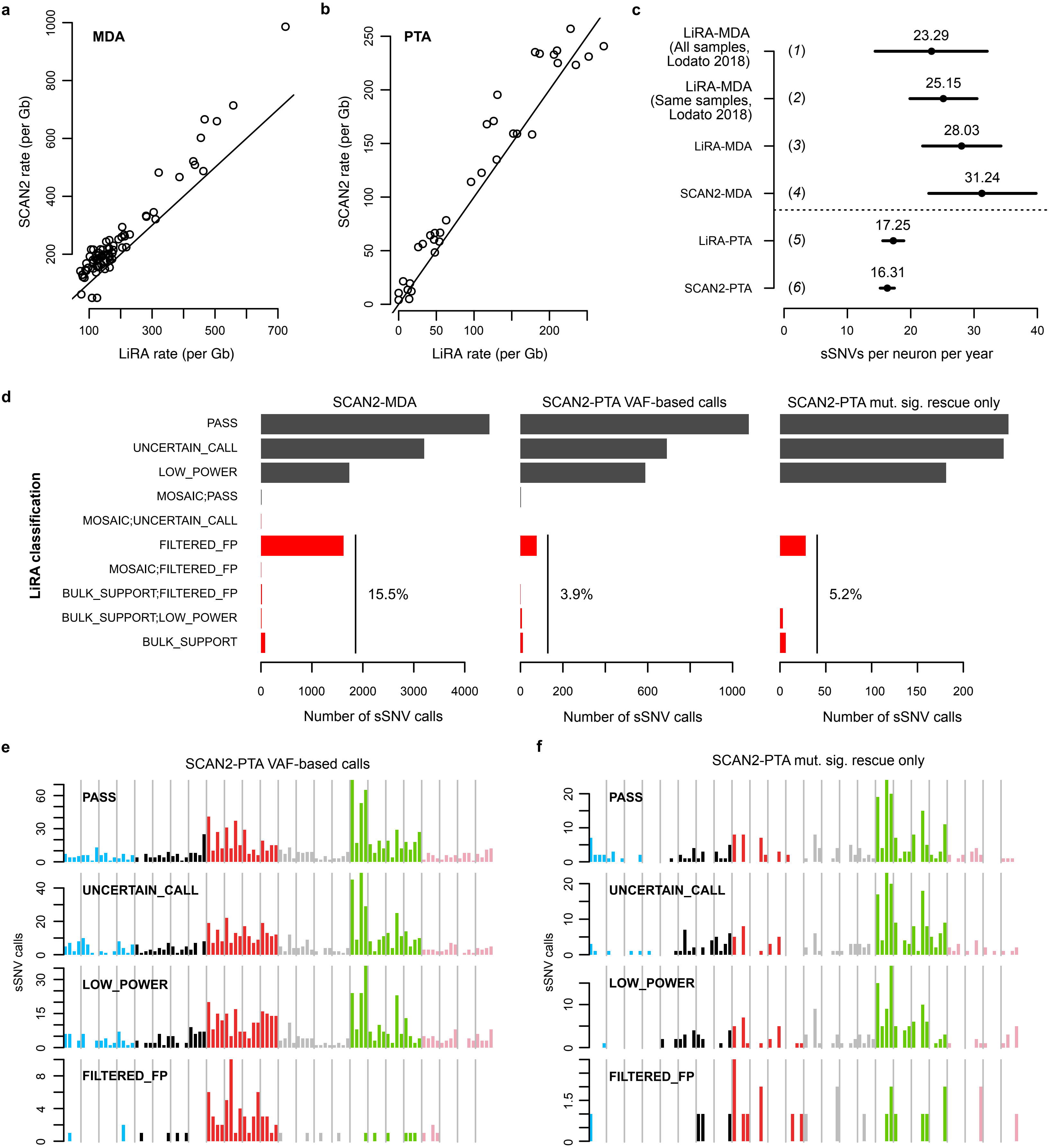

a
54 tissues from GTEx
Non-brain tissue
Brain tissue
Somatic SNV (obs/exp)
0.9
1.0
1.1
1.2
b
Somatic indel (obs/exp)
1.0
1.5
2.0
2
4
6
8
10
GTEx gene expression
Decile
e
98 tissues from Roadmap Epigenomics
Somatic SNVs
Somatic indels
Distal H3K27ac peaks
Significance: −log10(p−value)
0
1
2
3
4
Dorsolateral
Prefrontal
Cortex
Dorsolateral
Prefrontal
Cortex
f
Brain tissue
Non-brain tissue
Not significant
TSS proximal H3K27ac peaks
Significance: −log10(p−value)
Dorsolateral
Prefrontal
Cortex
Dorsolateral
Prefrontal
Cortex
0.6
0.8
1.0
1.2
1.4
0.5
1.0
1.5
2.0
2.5
Enrichment (or depletion) ratio (obs/exp)
Enrichment (or depletion) ratio (obs/exp)
c
High impact somatic mutations
0
20
40
60
80
Indels
SNVs
Extrapolated
Observed
d
sSNVs
Indels
Observed / expected
0.8
1.0
1.2
1.4
1.6
***
*
**
2
4
6
8
10
Conservation (phyloP 100)
Decile
g
Enhancers
h
Promoters
i
Open chromatin
(ATAC-seq)
j
DNA repair
hotspots
Observed / expected
0.5
1.0
1.5
2.0
2.5
3.0
3.5
***
**
*
***
**
**
***
*
***
**
***
*
***
*
astrocyte
neuron
microglia
oligo
astrocyte
neuron
microglia
oligo
astrocyte
neuron
microglia
oligo
astrocyte
neuron
microglia
oligo
GABA
GLU
OLIG
MGAS
GABA
GLU
OLIG
MGAS
SAR-seq
Repair-seq
SAR-seq
Repair-seq

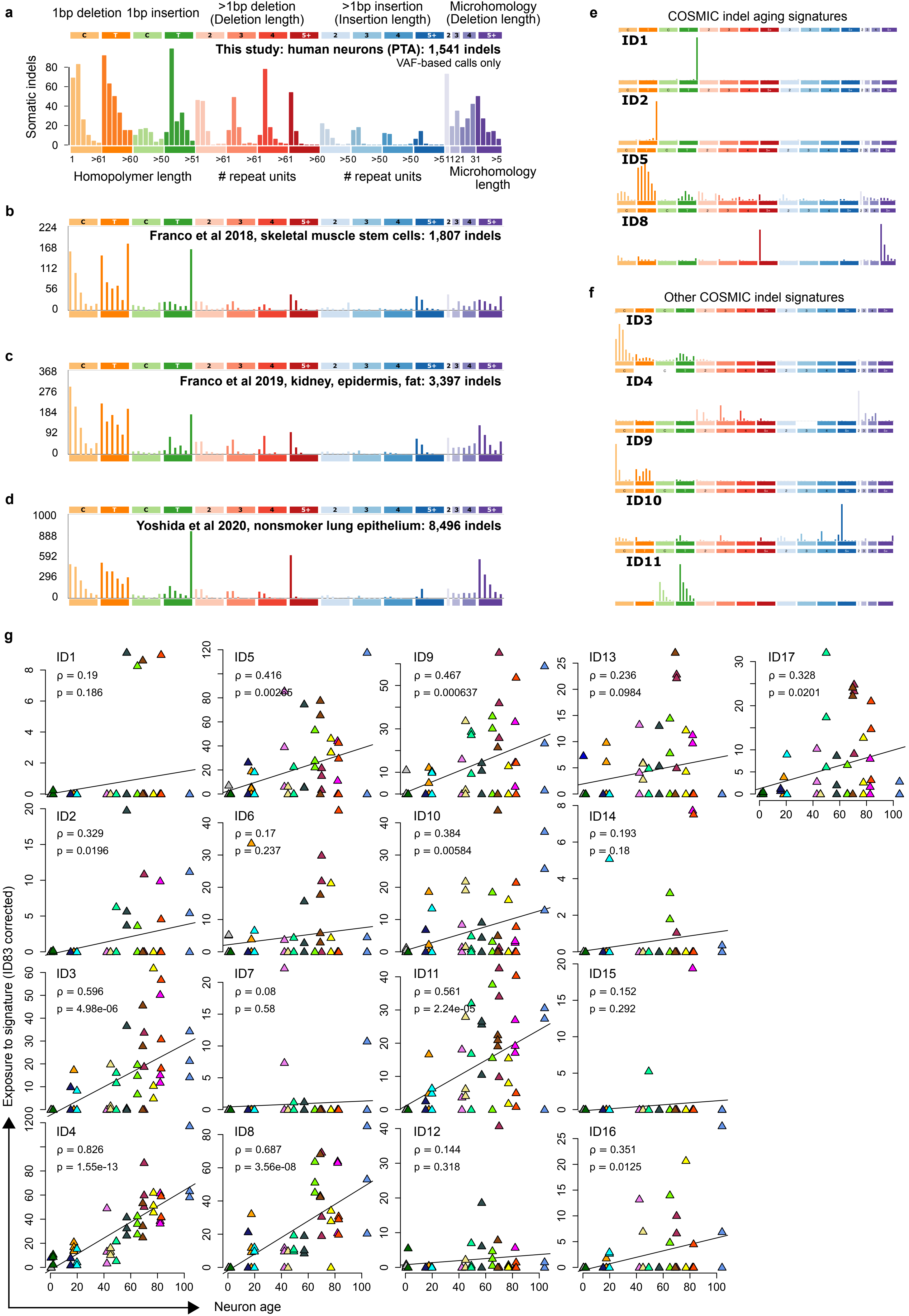

a
1278BA9-A
1278BA9-B
1278BA9-C
Allele balance
0.0
0.4
0.8
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
Chromosome
b
Allele balance
0.0 0.2 0.4 0.6 0.8 1.0
1278BA9−A
1278BA9−B
1278BA9−C
c
PTA
|1/2 - mean(AB)|
0.00
0.10
0.20
d
MDA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
Chromosome
e
AB correlation
1
0.8
0.6
0.4
0.2
0
4638-Neuron-6
4638-Neuron-4
4638-Neuron-5
4643-Neuron-3
4643-Neuron-4
4643-Neuron-6
UMB5572_D4
UMB5572_D2
UMB5572_D3
UMB5451_B2
UMB5451_B3
UMB5451_B5
UMB4976_E3
UMB4976_E1
UMB4976_E2
5871-Neuron-5
5871-Neuron-6
5219-Neuron-5
5219-Neuron-2
5219-Neuron-4
UMB5943_C5
UMB5943_C2
UMB5943_C4
UMB5666_F1
UMB5666_F2
UMB5666_F5
1278BA9-C
1278BA9-A
1278BA9-B
1465BA9-A
5817PFC-C
5817PFC-A
5817PFC-B
5871-Neuron-4
5559PFC-A
5559PFC-B
5559PFC-C
5087PFC-B
5087PFC-A
5087PFC-C
1465BA9-D
1465BA9-B
1465BA9-C
5823PFC-B
5823PFC-A
5823PFC-C
5657PFC-C
5657PFC-A
5657PFC-B
936PFC-B
936PFC-A
936PFC-C
PTA single cell sample

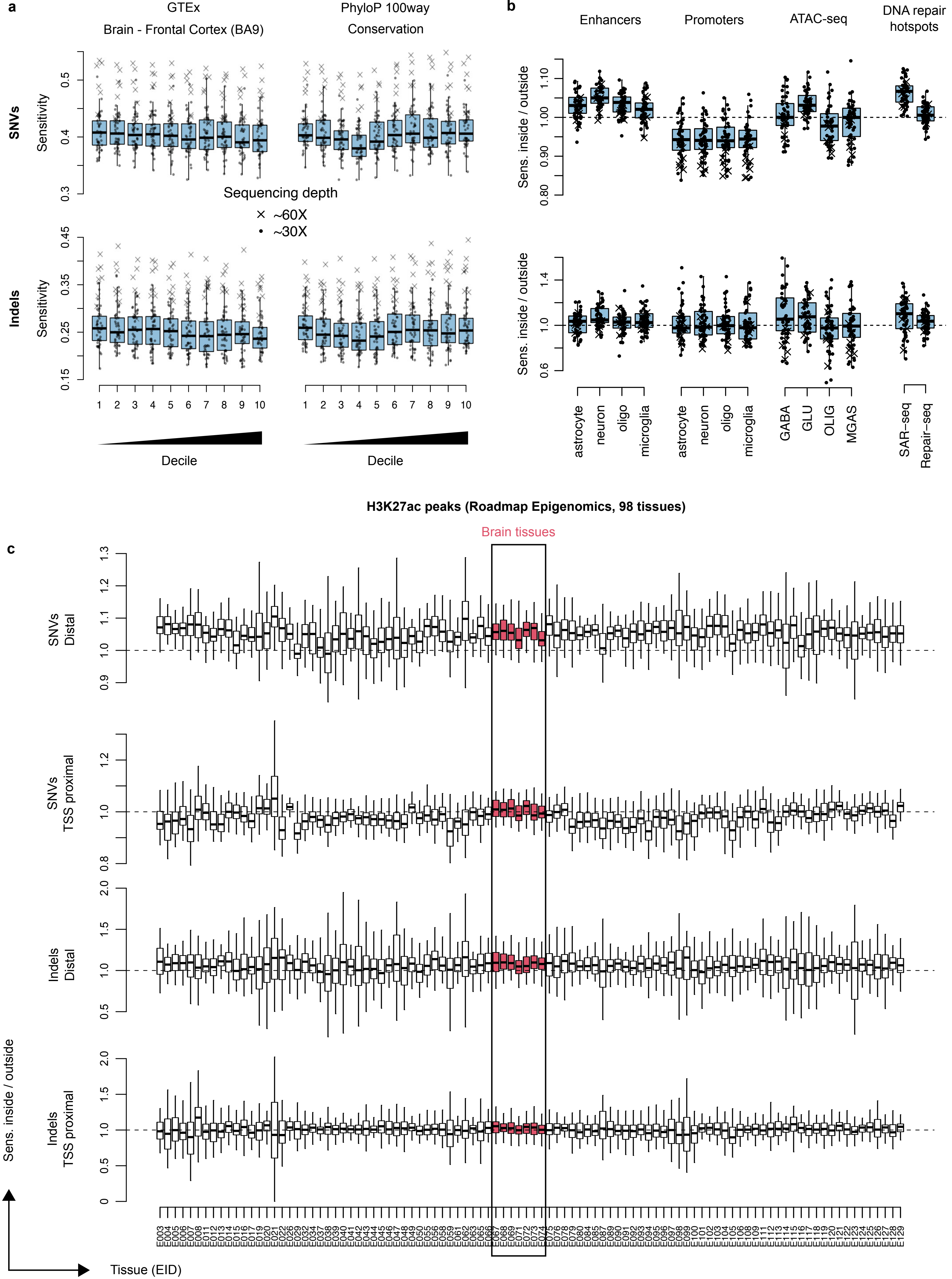

a
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
Spike-in SNVs per Mb
b
False discovery rate (%)
0
20
40
60
80
100
Spike-in SNVs per Mb
SCAN-SNV
SCAN2
c
Sensitivity
0.00
0.05
0.10
0.15
0.20
False positives per Mb
d
False discovery rate (%)
0.01
0.05
0.50
5.00
Mutation burden (sSNVs per Mb)
SCAN-SNV
SCAN2
Monovar
SCcaller (standard)
SCcaller (stringent)
e
Sensitivity relative to max performance
0.0
0.4
0.8
1.2
0
500
1500
# sSNVs used for signature construction
f
False discovery rate relative to max performance
0.0
0.4
0.8
1.2
0
500
1500
# sSNVs used for signature construction
g
SCAN2 SNV sensitivity
Regression (excl. PTAerr)
Slope: -0.035
P-value: 0.16
SBS88
SBS16
SBS12
SBS4
SBS9
SBS1
SBS54
SBS6
SBS2
SBS32
SBS3
SBS5
SBS11
SBS7b
SBS23
SBS19
SBS30
PTAerr
0.2
0.4
0.6
0.8
1.0
Cosine similarity to universal PTA error signature
h
SBS30
i
PTAerr

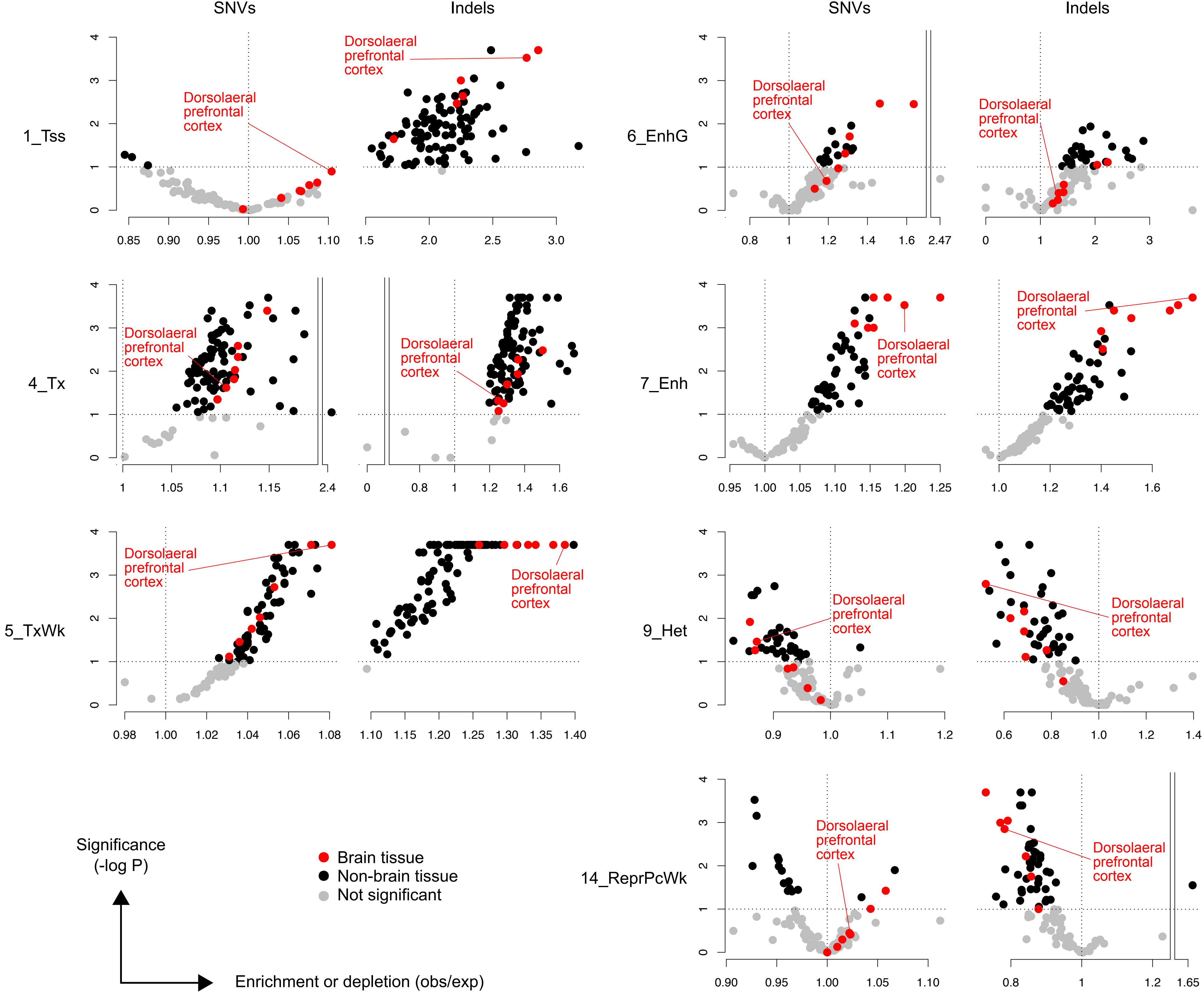

a
Untreated clone
LiRA
0
30
SCAN2 without mut. sig. rescue
0
20
50
SCAN2 rescued by signature
0
10
20
SCAN2 rejected by signature
0
20
50
b
Clonally supported SCAN2 calls (true positives)
0
20
50
Clonally unsupported SCAN2 calls (false positives)
0
5
15
c
Aristolochic acid (AAI) treated clone
LiRA
0
15000
SCAN2 without mut. sig. rescue
0
4000
SCAN2 rescued by signature
0
1500
SCAN2 rejected by signature
0
300
700

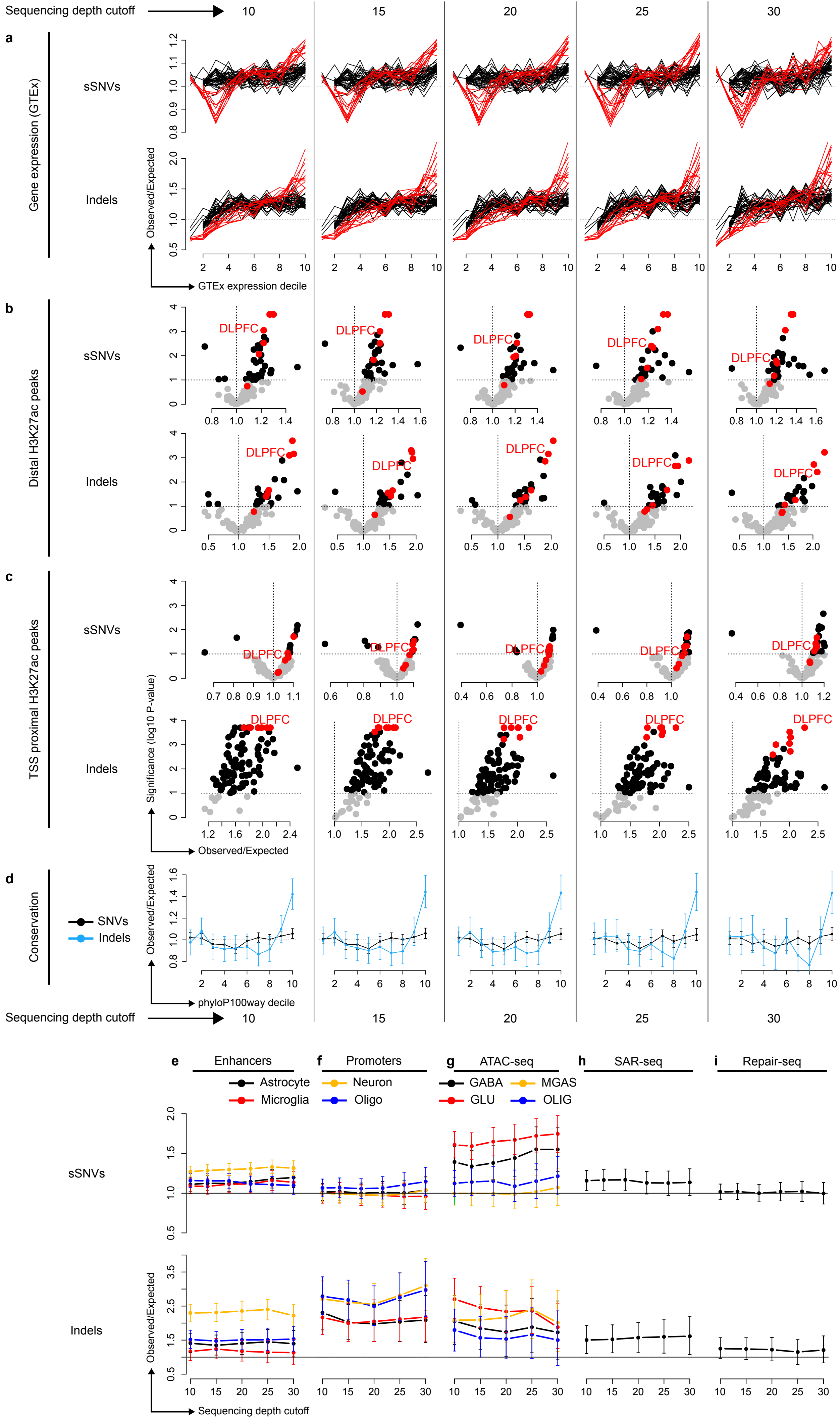

a
GATK
GATK w/cross-sample filter
SCAN–SNV
SCAN2 VAF-based only
SCAN2
Somatic indel calls per cell
10^4
10^2
1
b
False discovery rate
0
0.2
0.4
0.6
0.8
1
c
Sensitivity
Intermediate
High
Very high
Simulated mutation burden
d
All simulated indels
VAF-based SCAN2 calls
Number of indel spike-ins
0
500
1000
1500
2000
2500
-20
10
0
10
20
Indel length (<=0 = deletion)
e
VAF-based SCAN2 calls
Indel spike-in sensitivity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
f
VAF-based SCAN2 calls only
Final calls
Before cross–subject filter
Indel sensitivity
0.0
0.2
0.4
0.6
g
SCAN2
Indel ID83 mutation class
1:Del:C:0
1:Del:C:1
1:Del:C:2
1:Del:C:3
1:Del:C:4
1:Del:C:5
1:Del:T:0
1:Del:T:1
1:Del:T:2
1:Del:T:3
1:Del:T:4
1:Del:T:5
1:Ins:C:0
1:Ins:C:1
1:Ins:C:2
1:Ins:C:3
1:Ins:C:4
1:Ins:C:5
1:Ins:T:0
1:Ins:T:1
1:Ins:T:2
1:Ins:T:3
1:Ins:T:4
1:Ins:T:5
2:Del:R:0
2:Del:R:1
2:Del:R:2
2:Del:R:3
2:Del:R:4
2:Del:R:5
3:Del:R:0
3:Del:R:1
3:Del:R:2
3:Del:R:3
3:Del:R:4
3:Del:R:5
4:Del:R:0
4:Del:R:1
4:Del:R:2
4:Del:R:3
4:Del:R:4
4:Del:R:5
5:Del:R:0
5:Del:R:1
5:Del:R:2
5:Del:R:3
5:Del:R:4
5:Del:R:5
2:Ins:R:0
2:Ins:R:1
2:Ins:R:2
2:Ins:R:3
2:Ins:R:4
2:Ins:R:5
3:Ins:R:0
3:Ins:R:1
3:Ins:R:2
3:Ins:R:3
3:Ins:R:4
3:Ins:R:5
4:Ins:R:0
4:Ins:R:1
4:Ins:R:2
4:Ins:R:3
4:Ins:R:4
4:Ins:R:5
5:Ins:R:0
5:Ins:R:1
5:Ins:R:2
5:Ins:R:3
5:Ins:R:4
5:Ins:R:5
2:Del:M:1
3:Del:M:1
3:Del:M:2
4:Del:M:1
4:Del:M:2
4:Del:M:3
5:Del:M:1
5:Del:M:2
5:Del:M:3
5:Del:M:4
5:Del:M:5


a
MDA
SCAN2 rate (per Gb)
1000
800
600
400
200
100
300
500
700
LiRA rate (per Gb)
b
PTA
SCAN2 rate (per Gb)
250
200
150
100
50
0
0
50
100
200
LiRA rate (per Gb)
c
LiRA-MDA (All samples, Lodato 2018)
(1)
23.29
LiRA-MDA (Same samples, Lodato 2018)
(2)
25.15
LiRA-MDA
(3)
28.03
SCAN2-MDA
(4)
31.24
LiRA-PTA
(5)
17.25
SCAN2-PTA
(6)
16.31
0
10
20
30
40
sSNVs per neuron per year
d
LiRA classification
PASS
UNCERTAIN_CALL
LOW_POWER
MOSAIC;PASS
MOSAIC;UNCERTAIN_CALL
FILTERED_FP
MOSAIC;FILTERED_FP
BULK_SUPPORT;FILTERED_FP
BULK_SUPPORT;LOW_POWER
BULK_SUPPORT
SCAN2-MDA
15.5%
0
1000
2000
3000
4000
Number of sSNV calls
SCAN2-PTA VAF-based calls
3.9%
0
200
600
1000
Number of sSNV calls
SCAN2-PTA mut. sig. rescue only
5.2%
0
50
100
150
200
Number of sSNV calls
e
SCAN2-PTA VAF-based calls
sSNV calls
PASS
60
30
0
UNCERTAIN_CALL
40
20
0
LOW_POWER
30
15
0
FILTERED_FP
8
4
0
f
SCAN2-PTA mut. sig. rescue only
sSNV calls
PASS
20
10
0
UNCERTAIN_CALL
20
10
0
LOW_POWER
10
0
FILTERED_FP
1.5
0


a
1bp deletion
1bp insertion
>1bp deletion (Deletion length)
>1bp insertion (Insertion length)
Microhomology (Deletion length)
C
T
C
T
2
3
4
5+
2
3
4
5+
2
3
4
5+
This study: human neurons (PTA): 1,541 indels
VAF-based calls only
Somatic indels
0
20
40
60
80
1
>6
1
>6
0
>5
0
>5
1
>6
1
>6
1
>6
0
>5
0
>5
0
>5
1
1
2
1
3
1
>5
Homopolymer length
# repeat units
# repeat units
Microhomology length
b
Franco et al 2018, skeletal muscle stem cells: 1,807 indels
0
56
112
168
224
c
Franco et al 2019, kidney, epidermis, fat: 3,397 indels
0
92
184
276
368
d
Yoshida et al 2020, nonsmoker lung epithelium: 8,496 indels
0
296
592
888
1000
e
COSMIC indel aging signatures
ID1
ID2
ID5
ID8
f
Other COSMIC indel signatures
ID3
ID4
ID9
ID10
ID11
g
ID1
ρ = 0.19
p = 0.186
ID2
ρ = 0.329
p = 0.0196
ID3
ρ = 0.596
p = 4.98e-06
ID4
ρ = 0.826
p = 1.55e-13
ID5
ρ = 0.416
p = 0.00265
ID6
ρ = 0.17
p = 0.237
ID7
ρ = 0.08
p = 0.58
ID8
ρ = 0.687
p = 3.56e-08
ID9
ρ = 0.467
p = 0.000637
ID10
ρ = 0.384
p = 0.00584
ID11
ρ = 0.561
p = 2.24e-05
ID12
ρ = 0.144
p = 0.318
ID13
ρ = 0.236
p = 0.0984
ID14
ρ = 0.193
p = 0.18
ID15
ρ = 0.152
p = 0.292
ID16
ρ = 0.351
p = 0.0125
ID17
ρ = 0.328
p = 0.0201
Exposure to signature (ID83 corrected)
Neuron age
0
20
40
60
80
100


a
GTEx
Brain - Frontal Cortex (BA9)
PhyloP 100way
Conservation
SNVs
Sensitivity
0.3
0.4
0.5
Sequencing depth
× ~60X
• ~30X
Indels
Sensitivity
0.15
0.25
0.35
0.45
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
Decile
Decile
b
Enhancers
Promoters
ATAC-seq
DNA repair
hotspots
Sens. inside / outside
0.80
0.90
1.00
1.10
Sens. inside / outside
0.6
1.0
1.4
astrocyte
neuron
oligo
microglia
astrocyte
neuron
oligo
microglia
GABA
GLU
OLIG
MGAS
SAR−seq
Repair−seq
H3K27ac peaks (Roadmap Epigenomics, 98 tissues)
Brain tissues
c
SNVs
Distal
0.9
1.0
1.1
1.2
1.3
SNVs
TSS proximal
0.8
1.0
1.2
Indels
Distal
0.5
1.0
1.5
2.0
Indels
TSS proximal
0
0.5
1.0
1.5
2.0
Sens. inside / outside
Tissue (EID)
E003 E004 E005 E006 E007 E008 E011 E012 E013 E014 E015 E016 E017 E019 E020 E021 E022 E026 E029 E032 E034 E037 E038 E039 E040 E041 E042 E043 E044 E045 E046 E047 E048 E049 E050 E055 E056 E058 E059 E061 E062 E063 E065 E066 E067 E068 E069 E071 E072 E073 E074 E075 E076 E078 E079 E080 E084 E085 E087 E089 E090 E091 E092 E093 E094 E095 E096 E097 E098 E099 E100 E101 E102 E103 E104 E105 E106 E108 E109 E110 E111 E112 E113 E114 E115 E116 E117 E118 E119 E120 E121 E122 E123 E124 E125 E126 E127 E128 E129


SNVs
Indels
SNVs
Indels
1_Tss
6_EnhG
4_Tx
7_Enh
5_TxWk
9_Het
14_ReprPcWk
Dorsolaeral prefrontal cortex
0
1
2
3
4
0.85
0.90
0.95
1.00
1.05
1.10
1.5
2.0
2.5
3.0
0.8
1
1.2
1.4
1.6
2.47
0
1
2
3
1
1.05
1.1
1.15
2.4
0
0.8
1
1.2
1.4
1.6
0.95
1.00
1.05
1.10
1.15
1.20
1.25
1.0
1.2
1.4
1.6
0.98
1.00
1.02
1.04
1.06
1.08
1.10
1.15
1.20
1.25
1.30
1.35
1.40
0.9
1.0
1.1
1.2
0.6
0.8
1.0
1.2
1.4
0.90
0.95
1.00
1.05
1.10
0.8
1
1.2
1.65
Significance (-log P)
Enrichment or depletion (obs/exp)
● Brain tissue
● Non-brain tissue
● Not significant


Sequencing depth cutoff
10
15
20
25
30
a
Gene expression (GTEx)
sSNVs
Indels
Observed/Expected
0.8
0.9
1.0
1.1
1.2
0.5
1.0
1.5
2.0
2
4
6
8
10
GTEx expression decile
b
Distal H3K27ac peaks
sSNVs
Indels
DLPFC
0
1
2
3
4
0.8
1.0
1.2
1.4
1.6
0.5
1.0
1.5
2.0
c
TSS proximal H3K27ac peaks
sSNVs
Indels
DLPFC
Significance (log10 P-value)
0.4
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.6
2.0
2.4
2.5
Observed/Expected
d
Conservation
SNVs
Indels
Observed/Expected
0.8
1.0
1.2
1.4
1.6
phyloP100way decile
Sequencing depth cutoff
10
15
20
25
30
e
Enhancers
f
Promoters
g
ATAC-seq
h
SAR-seq
i
Repair-seq
Astrocyte
Microglia
Neuron
Oligo
GABA
GLU
MGAS
OLIG
sSNVs
Indels
Observed/Expected
0.5
1.0
1.5
2.0
2.5
3.5
10
15
20
25
30
Sequencing depth cutoff

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); i.e., 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 the strongest correlation.

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: FDR = *F* / (*F* + x*S*). SCcaller standard uses a calling threshold of α = 0.05 while stringent calling uses α = 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 (**i**).

Extended Data Fig. 3. Mutation spectra of SCAN2 and LiRA calls on kindred mouse ESC cells.
a-b. SBS96 signatures 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.** Signatures for SNVs called in the 4 single cells taken from an aristolochic acid (AAI)-treated clone.

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 rates 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. Comparison of SCAN2 and LiRA sSNV calls on human neurons. Single human neurons were previously 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. Somatic indels mutation spectra in human neurons and other cells.

a. Spectrum of 1541 indels from PTA neurons from this study, same as **Figure 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 **Figure 4**. Note that y-axes are not the same scale.

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. Patterns of mutation enrichment persist at increasing sequencing depth thresholds.

Analyses presented in **Figure 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.

Supplementary Note

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

Lovelace J. Luquette, Michael B. Miller, Zinan Zhou, Craig L. Bohrson, Yifan Zhao, Hu Jin, Doga Gulhan, Javier Ganz, Sara Bizzotto, Samantha Kirkham, Tino Hochepied, Claude Libert, Alon Galor, Junho Kim, Michael A. Lodato, Juan I. Garaycoechea, Charles Gawad, Jay West, Christopher A. Walsh and Peter J. Park.

Large somatic copy number alteration analysis

Ginkgo[40] was first applied to produce normalized read counts for bulks and PTA single cells using 100 kb variable-sized binning. First, 1 MB bins were created by merging non-overlapping runs of 10 100 kb bins. Next, large somatic CNA candidates on chromosomes 1-22 were defined as runs of 5 or more windows i (corresponding to ~5 MB) with read depth ratio $S_{j,i}/B_i < 0.6$ or > 1.4, where $S_{j,i}$ denotes the normalized read depth in window i in single cell j and B_i is the same normalized window in the matched bulk sample. This CNA calling procedure is crude and only intended to recover very large (>5 MB) CNAs; however, these parameters successfully recovered male X chromosomes and female Y chromosomes in bulk and the large deletions observed in the PTA-amplified neuron 5823PFC-B and 5871-Neuron-4 (**Supplementary Figure 1**). All 4 PTA cells from subject 1465 had unusually noisy profiles and may reflect poor tissue quality and thus were excluded from CNA analysis.

False discovery rate estimation for MDA and PTA

Estimated FDR curves shown in **Supplementary Figure 5** were parameterized by

$$\mathrm{FDR} = \frac{\text{FP rate per Mb}}{\text{FP rate per Mb} + \text{Sensitivity} \times \text{Mutations per Mb}}$$

Parameters used were: PTA with GATK (ref. 16), FP rate per Mb = 0.9, sensitivity = 0.8; PTA with SCAN2 (mutation-signature and VAF-based calls) FP rate per Mb = 0.0143, sensitivity = 0.458 (taken from simulation experiments, see *Synthetic diploid X-chromosome simulations*). To compute the best-case scenario for MDA, we assumed that all artifacts caused by single stranded dropout would be erroneously identified as true SNVs and that these would be the only source of FPs. The number of single-stranded dropout artifacts in MDA was estimated by the excess number of sSNV calls per hemizygous X chromosome (15.9 sSNVs). Excess MDA calls were calculated per individual as $\mathrm{median}(\#\text{ corrected MDA calls}) - \mathrm{median}(\#\text{ corrected PTA calls})$. To convert to FPs per diploid megabase, the excess rate is first doubled and then divided by 152,231,524 bp, the size of chromosome X after removing pseudoautosomal regions. This yielded a rate of 0.21 FPs per Mb, which was applied to the whole genome. Finally, we assume these FPs are called with the same sensitivity as true mutations since they do not cause allelic imbalance or discordantly linked reads. Thus, there was no need to provide a sensitivity parameter for the best-case MDA scenario since it would cancel out in the above equation.

Deriving the universal PTA artifact spectrum

The SCAN2 package provides universal PTA artifact spectra for both SBS96 and ID83 signature formats for calling somatic SNVs and indels, respectively. Each universal PTA artifact spectrum (separately for SNVs and indels) was derived in 2 steps (technical details are provided in the next paragraph). First, two sets of mutations enriched for artifacts were extracted from each cell (**Supplementary Figure 3a**): (1) $S_{\text{X artifact}}$ from X chromosomes (male samples only) and (2) $S_{\text{Autosomal artifact}}$ from autosomal mutation candidates with VAFs consistent with expectation for pre-amplification artifacts, as determined by the local allele balance. $S_{\text{Autosomal artifact}}$ was added because $S_{\text{X artifact}}$

consisted of only 190 likely artifacts, which may be insufficient to produce a high quality mutation spectrum. Second, *de novo* signature extraction was performed on $S_{\text{X artifact}}$, $S_{\text{Autosomal artifact}}$ and an additional set S_{PASS} of high quality mutations produced by VAF-based calling (**Supplementary Fig. 3b**). The high quality mutation set provides the true mutational signature, helping to prevent true mutations in $S_{\text{X artifact}}$ or $S_{\text{Autosomal artifact}}$ from being assigned to the artifact signature. For both SNVs and indels, *de novo* signature extraction produced N=2 signatures, as expected: one corresponding to S_{PASS} and a second corresponding to the PTA high-VAF artifact process, which became the universal PTA artifact spectrum (**Supplementary Fig. 3c**). Estimated exposures to the true and artifact spectra confirmed that the two artifact sets were highly enriched for artifacts, contrasting with the high-quality set (**Supplementary Fig. 3d** for SNVs). The similarity between the sSNV PTA universal artifact signature and the MDA artifact C>T signature is notable and provides evidence that the signature is unlikely to be an overfit to this dataset. The indel PTA universal artifact signature is characterized by 1 bp insertions in homopolymers and 1 bp T:A deletions in homopolymers (**Supplementary Fig. 3e**).

In more detail, X chromosome artifacts were identified from candidate mutations produced by GATK HaplotypeCaller (as described in *Comparison of MDA and PTA somatic mutation calls*) by requiring the candidate to: (1) occur in the non-pseudoautosomal X regions, (2) have total sequencing depth >= median(sequencing depth) of the X chromosome, (3) be supported by at least 6 alternate reads, and (4) have 35% <= VAF <= 75%. Autosomal artifacts were identified by the SCAN2 allele balance consistency (ABC, P_{true}) and pre-amplification test (P_{artifact}) *P*-values (see ref. 14). Briefly, large ABC *P*-values indicate that the candidate mutation's VAF is consistent with the locally estimated allele balance, as should be the case for a true mutation. Large pre-amplification *P*-values indicate that the candidate's VAF is consistent with that expected for an early-occuring artifact. These two cases are not mutually exclusive, particularly for mutations at 33% VAF in a region of 2:1 imbalance. Autosomal mutation candidates which fail the pre-amplification test, pass all other SCAN2 tests and for which $P_{\text{amplification artifact}} > P_{\text{ABC}}$ were selected as autosomal artifacts. S_{PASS} is the set of mutations called by SCAN2 using VAF-based calling with the stringent calling–parameter `–target.fdr`=0.01. *De novo* signature extraction was performed by `SigProfiler`[43] version 2.5.1.7, as used in other *de novo* extractions. Signature channels with values < 10^{-4} were replaced by 10^{-5} to prevent channels with extreme weights.

Multi-sample test for mutation signature homogeneity

The *P*-values shown in **Supplementary Figure 11** were automatically generated during signature-based rescue (`scan2 rescue`). The batch test is computed as follows. Let $\vec{T}$ denote the mutation spectrum (e.g., SBS96 and ID83 spectra are used in this work, but any spectrum can be used in principle) of calls combined across all cells in a batch. For single cell *i*, let N_i denote the number of somatic mutations called and $\vec{S}_i$ denote the mutation spectrum of those calls. We determine whether $\vec{S}_i$ differs from $\vec{T}$ with the following procedure. 1 million random spectra $R_i^{(j)}$, j = { 1 .. 10^6 } of size N_i are drawn from a multinomial distribution with probability vector $\vec{T}$ and their log-likelihoods $L_i^{(j)}$ under the same multinomial model are computed. I.e.,

$$R_i^{(j)} \sim \text{Multinomial}\left(N_i;\ \vec{T}\right), \quad L_i^{(j)} = \log L\left(R_i^{(j)} \mid \vec{T}\right).$$

The log-likelihood L_i under the same multinomial model is also computed for the single cell spectrum $\vec{S}_i$. The probability that single cell *i* derives from the group-wide spectrum is then

$$P_i = \frac{\left|\left\{j:\ R_i^{(j)} < L_i\right\}\right|}{10^6}.$$

Signature-based somatic mutation calling with SCAN2

First, a set of high quality somatic mutation (either sSNV or indel) calls is produced by VAF-based calling with a stringent target FDR of 1%. The true mutation spectrum is produced by combining calls from all cells supplied to SCAN2 into a single, raw mutation spectrum (currently SBS96 format for sSNV analysis and the ID83 format for indel analysis; however, other signature formats can be used in principle). If a batch of several single cells are available that can reasonably be expected to contain the same mutation signatures, then we recommend combining the calls across the batch to increase the accuracy of true signature estimation. We provide a test to determine whether a batch of cells violates this mutation signature homogeneity (described in *Multi-sample test for mutation signature homogeneity*). Next, only mutations satisfying the following criteria are eligible for rescue: (1) the mutation must pass all SCAN2 filters except the pre-amplification artifact filter and (2) the mutation must have estimated pre-amplification and amplification FDR no greater than 0.5 (see ref. 14, *Determining p-value cutoffs* for details on these artifact statistical tests and the FDR procedure). The fraction of mutations generated by the true spectrum and universal PTA artifact spectrum (described below) are estimated for each single cell by least squares fitting. Weights are computed for each cell *i* and rejected mutation candidate *j* using a likelihood ratio

$$W_{i,j} = \frac{P\left(\text{Mut. channel}\left(\text{mut}_{i,j}\right) \mid \text{True spectrum}\right) P(\text{True spectrum} \mid \text{cell}_i)}{P\left(\text{Mut. channel}\left(\text{mut}_{i,j}\right) \mid \text{Artifact spectrum}\right) P(\text{Artifact spectrum} \mid \text{cell}_i)},$$

where $P\left(\text{Mut. channel}\left(\text{mut}_{i,j}\right) \mid \text{True spectrum}\right)$ is the component of the true mutation spectrum corresponding to the mutation type (e.g., trinucleotide context and base change when using SBS96) of $\text{mut}_{i,j}$ and $P(\text{True spectrum} \mid \text{cell}_i)$ is cell *i*'s estimated fraction of mutations generated by the true mutation signature. The same meanings apply to the artifact spectrum. Therefore, $W_{i,j} > 1$ indicates lower likelihood of $\text{mut}_{i,j}$ being produced by the artifact process while $W_{i,j} < 1$ indicates higher likelihood. The weight is used to adjust a previously described FDR heuristic[14] that estimates the ratio of true mutations N_T and artifacts N_A among candidate mutations with similar VAF and sequencing depth as mut$_{i,j}$. This produces a multi-sample adjusted, Phred-scaled quality score $Q'_{i,j}$:

$$Q'_{i,j} = -10 \log_{10} \mathrm{FDR}_{i,j} = -\,10 \log_{10} \left\{ \frac{\alpha_{i,j}}{\alpha_{i,j} + \beta_{i,j} \cdot \frac{N_{T,i,j}}{N_{A,i,j}} \cdot W_{i,j}} \right\},$$

where $\alpha_{i,j}$ and $\beta_{i,j}$ are the type I error rate and power for mut$_{i,j}$ estimated by the pre-amplification artifact model used by SCAN-SNV (ref. 14 provides more details on this model). Finally, the rejected candidate mut$_{i,j}$ is accepted if it was previously rejected only by the pre-amplification artifact model and $Q'_{i,j} > 20$, corresponding to a desired FDR of 1%. This threshold can be set by the user.

Estimation of genome-wide somatic mutation burden

Genome-wide somatic mutation burdens are computed by approximating somatic mutation recall with germline mutation sensitivity at a select subset of heterozygous germline variants. VAF-based, not mutation signature-rescued, somatic calls are used for this extrapolation because the signature-rescue process further complicates the somatic recall estimation process. SCAN2 estimates genome-wide mutation burden automatically during the `call_mutation` pipeline.

For both mutation types (SNVs and indels), the germline variant set begins as the full set of phased (by SHAPEIT2) germline variants detected in the matched bulk. Because the distance between any candidate somatic mutation and the nearest hSNP affects the accuracy of the spatial AB model and thus calling, it is critical to ensure that the population of germline variants is similarly distant from training hSNPs as the somatic candidate mutation population. Importance sampling is used to subsample the germline variants to match the distance-to-nearest-hSNP distribution of the somatic candidate set.

Each germline variant is then individually analyzed using a leave-1-out approach. Allele balance at the variant is predicted by the allele balance model with the variant removed from the training set (if it was used). The FDR heuristic N$_T$/N$_A$, which provides a prior estimate on the ratio of true mutations and artifacts, is partially recomputed since full recomputation is computationally intensive. The partial recomputation adds 1 to the estimated number of total true mutations N_T (to account for the left-out germline variant, which is now viewed as a true somatic mutation) and to the number of somatic candidates at a given VAF bin S_i (to treat the germline variant as if it were among the somatic candidate set). The VAF range [0,1] is binned into 20 equally sized bins of 0.05 each. Full recomputation would involve recomputing N_T using the method in ref. 14, *Estimating artifact prevalence*. Following the notation in ref. 14, the partially recomputed heuristic is:

$$N_{T,\mathrm{i}} = \max\left\{(N_T + 1)\frac{H_i}{H}, 0.1\right\}, \qquad N_{\mathrm{A},i} = \max\{(S_i + 1) - N_{T,\mathrm{i}}, 0.1\},$$

where N_T is an upper bound on the number of true somatic mutations in the candidate set; $N_{T,\mathrm{i}}$ is the same upper bound within VAF bin i; H_i is the number of heterozygous germline variants in VAF bin i and H is the total number of heterozygous germline variants; and S_i is the number of somatic candidate mutations in VAF bin i. Note that all of the above calculations are applied to candidate mutations and germline variants with the same sequencing depth; each depth is analyzed separately and produces a different N_T/N_A estimate. Finally, the germline variant is called using all somatic filters (which incorporates allele balance estimates and N_T/N_A) except for dbSNP exclusion and lack of supporting reads in bulk.

Analysis is then restricted to the subset of the genome with single cell sequencing depth between Q_{25} and Q_{75}, the 25th and 75th percentiles of sequencing depth of the germline variants and the user-specified minimum bulk depth requirement (by default, >10). The fraction f_h of passing germline variants in this region serves as an estimate of somatic sensitivity. The rate of somatic mutations per haploid gigabase is then

$$R_{\mathrm{Gb}} = \frac{N_{\mathrm{somatic}}/f_h}{2C},$$

where C is the number of diploid bases passing the single cell and bulk depth requirements above. C is collected by running GATK DepthOfCoverage on the single cell and matched bulk, which outputs single cell and bulk read depth at every position in the reference genome. C is the number of positions with single cell depth between Q_{25} and Q_{75} and bulk depth greater than the user-set minimum (by default, >10). The factor of 2 converts diploid bases to haploid bases. The final extrapolated value is $G \cdot R_{\mathrm{Gb}}$, where G is the genome size in haploid gigabases; for **Figure 4a**, G=5.845 corresponds to the size of GRCh37d5 chromosomes 1-22 and matches ref. 6; for synthetic diploid simulations, G=0.3044, corresponding to approximately twice the size of the haploid, non-pseudoautosomal region of chromosome X in GRCh37d5. **Supplementary Figure 7** provides an assessment of the accuracy of this estimate in simulated data with known mutation burdens.

Somatic indel detection with SCAN2

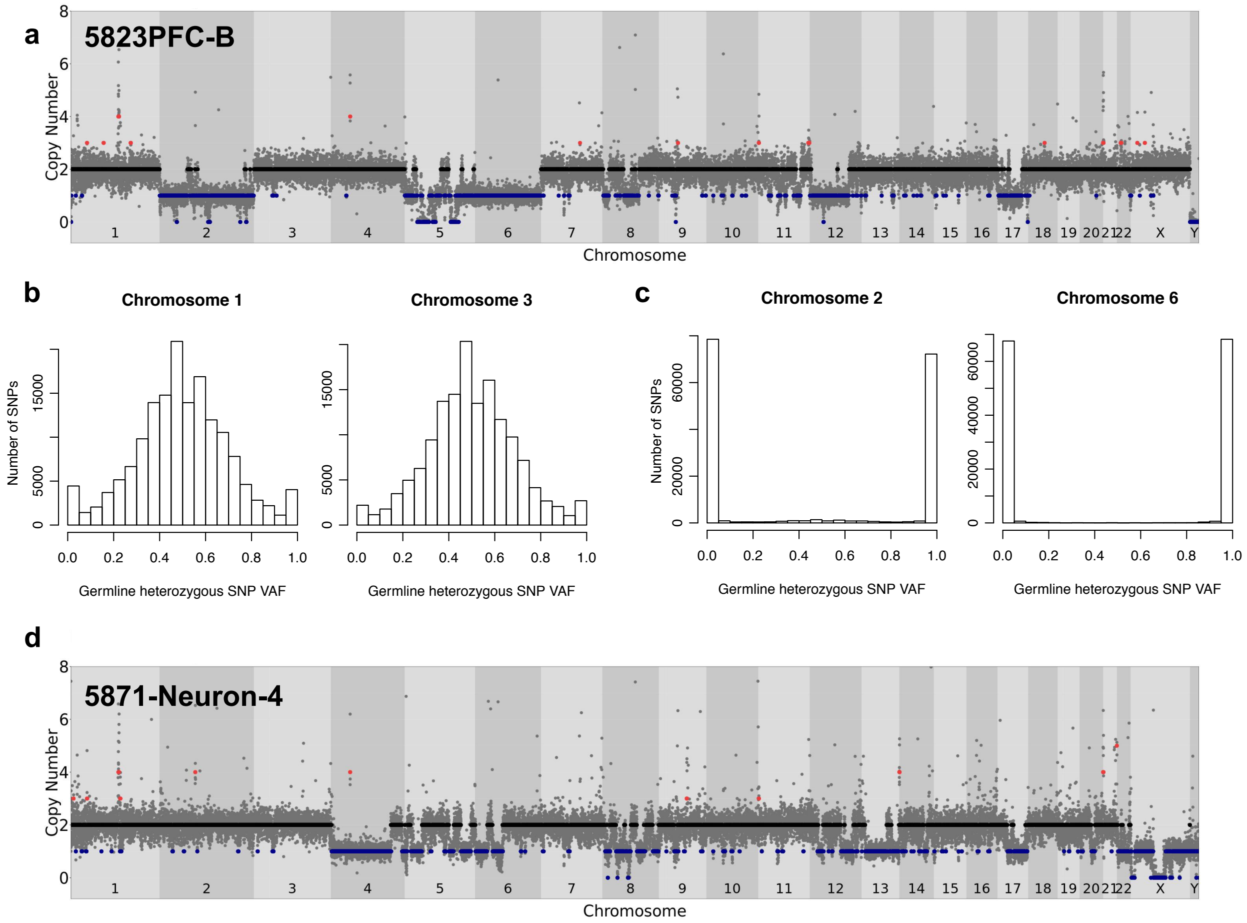
Counts of reference and alternate supporting reads at indels are generated in the same way as for sSNVs. As for sSNVs, BAMs from all single cell and bulk samples from a single donor are provided jointly to GATK. GATK was chosen to provide these read counts over alternatives (such as samtools) due to GATK's local reassembly that helps to prevent artifactual read pileups near indels. Only GATK's reference and alternate read counts at each locus are used—other genotyping scores and metrics produced by GATK are ignored. Each locus with non-reference read support (somatic indel candidate) is assessed by all tests and filters applied to somatic SNVs in VAF-based mode and an additional single-cell depth requirement of 10 reads. Notably, the allele balance model applied to candidate somatic indels is not built using germline indels; rather, the same model trained on germline hSNPs and applied to sSNVs is used for indel calling. Somatic indels passed by this process are then filtered using the cross-sample site list by requiring either: (1) reads supporting the somatic indel exist only in single cells from one individual or (2) no single cell other than the mutation-harboring cell contains more than 2 supporting reads, regardless of the number of cells and subjects in which these indel-supporting reads appear. The cross-sample list is generated by running GATK HaplotypeCaller (with the same parameters as in indel discovery) jointly on whole-genome amplified single cells from at least two individuals, including the one being analyzed. The cross-sample filter is applied both to mutations called in VAF-based mode and signature-based recovery. Signature-based rescue of indels, as described in *Signature-based somatic mutation calling with SCAN2*, uses the 83-channel indel format ID83 produced by SigProfilerMatrixGenerator[41].

VAF parsimony phasing

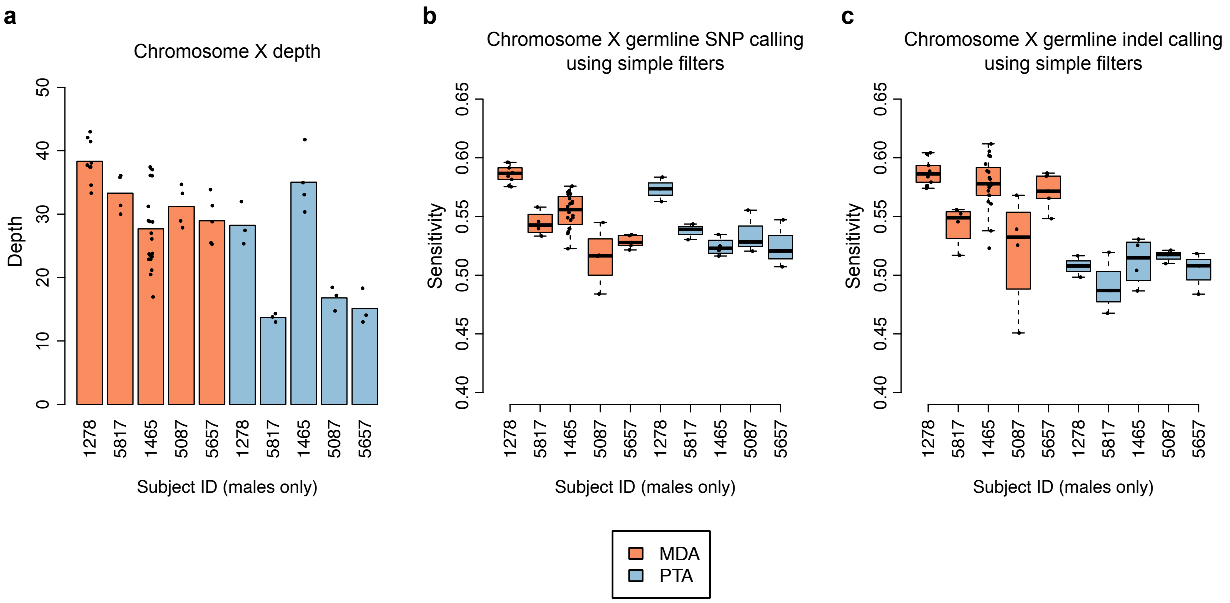
For crossbred mouse ESCs, population-based reference panel phasing would likely lead to near-perfect phasing (e.g. by using the *Mus musculus* and *Mus spretus* genomes as reference haplotypes). We therefore implemented an alternative phasing approach using VAF-parsimony (`--parsimony-phasing`): the phase of hSNP *i* was set to the same phase as hSNP *i*-1 if VAF_i was closer to VAF_{i-1}; otherwise hSNP *i* was assigned the opposite phase. This approach creates many phasing errors in balanced genomic regions because each allele will have VAF near 1/2; however, for the same reason, these phasing errors have a small impact on AB model accuracy. However, in regions with high levels of imbalance, VAFs will be near 0 for one allele and 1 for the other, making it clear on which allele the hSNP truly resides.

Removal of signature B from MDA samples

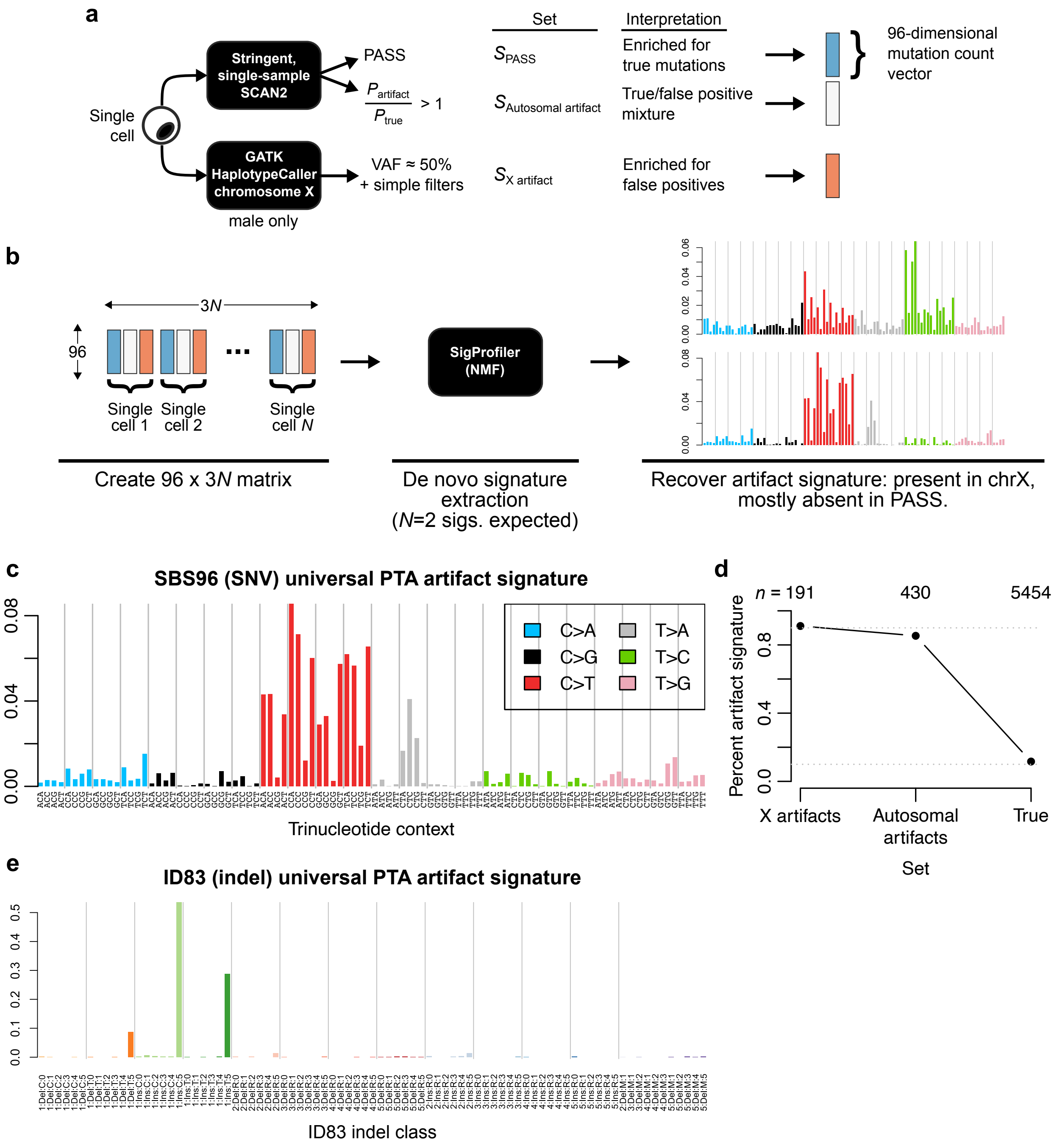
Signature B levels in MDA samples were measured by de novo signature extraction from the combined set of VAF-based sSNV calls from 128 PTA and MDA neurons using `SigProfiler` version 2.5.1.7. Three signatures were discovered, with one nearly identical to signature B[6] (cosine similarity=0.996). Removal of signature B as shown in **Supplementary Figure 9** was achieved by multiplying the total genome burden predictions for each cell by the fraction of mutations assigned to the two non-signature B signatures.



Supplementary Figure 1. PTA-amplified neurons with large-scale copy number changes **a.** Neuron B from subject 5823 shows single copy loss over the majority of chromosomes 2, 5, 6, 12 and 17. **b.** Variant allele fractions (VAF) for heterozygous germline SNPs on chromosomes 1 and 3 show the expected VAF variance for successfully amplified chromosomes. **c.** Same as (b) for chromosomes 2 and 6, which show a loss over the majority of each chromosome. VAF values at 0 and 1 are consistent with the complete loss of a single haplotype, ruling out the possibility that both alleles were present and amplified but to a lower level than other chromosomes. However, whether the neuron truly contained several single copy loss or if the apparent loss resulted from localized amplification failures of one haplotype cannot be determined. **d.** Copy number profile for Neuron-5871-4.

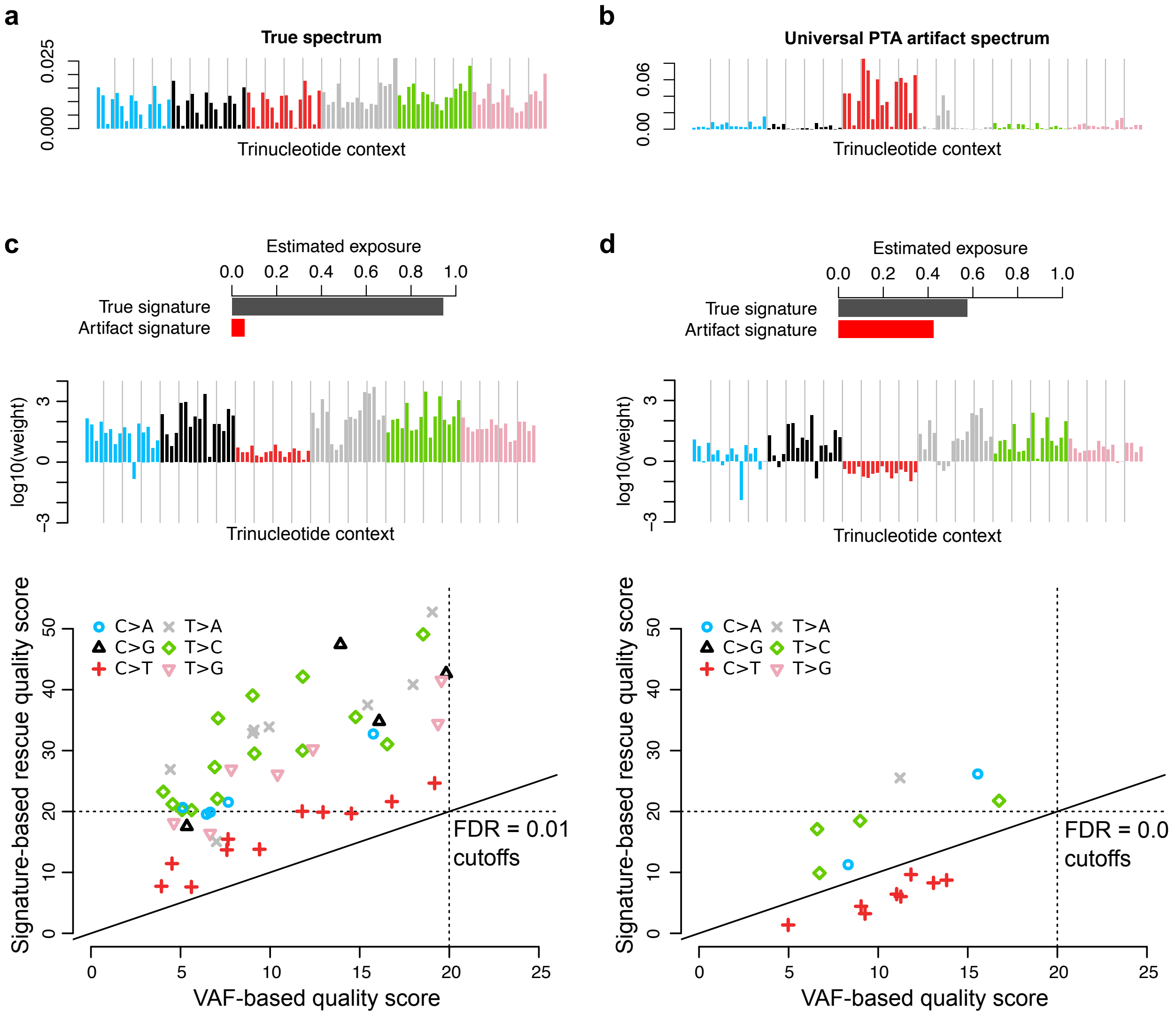


Supplementary Figure 2. Simple somatic mutation calling on male chromosome X. **a.** Mean sequencing depth per cell (points) and averaged over all cells per individual (bar). PTA cells for subjects 1278 and 1465 were sequenced to ~60X total depth while other PTA cells were sequenced to ~30X. Chromosome X in males should be covered at approximately half of the genome-wide mean sequencing depth due to hemizygosity. **b.** Sensitivity for germline SNPs using somatic SNV calling criteria (depth and allele fraction filters). Germline SNP sensitivity provides an estimate for somatic SNV sensitivity. **c.** Same as (b) for germline indels. Boxplot whiskers, the furthest outlier <=1.5 times the interquartile range from the box; box, 25th and 75th percentiles; centre bar, median.; *n*=16 PTA neurons and *n*=39 MDA neurons.

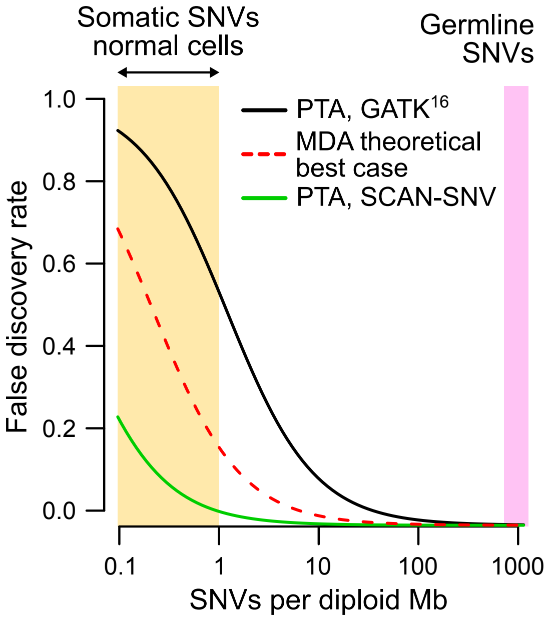


Supplementary Figure 3. The universal PTA SNV and indel artifact signatures.
a. 3 sets of SNVs and likely artifacts were constructed for each male single cell. PASS autosomal SNVs using stringent calling filters are highly depleted for artifacts whereas rejected candidate SNVs are highly enriched for early, high-VAF PTA artifacts. Rejected candidate SNVs are defined as those with either $P_{\text{artifact}}/P_{\text{true}} > 1$ (see ref. 14 for information on the models corresponding to these *P*-values) or chromosome X sites in the non-pseudoautosomal regions with ~50% VAF in male samples. **b.** An SBS96 mutation count matrix is constructed for de novo signature extraction using 3 separate entries for each male single cell (not shown: female cells are also used but have no X chromosome component). *De novo* signature extraction

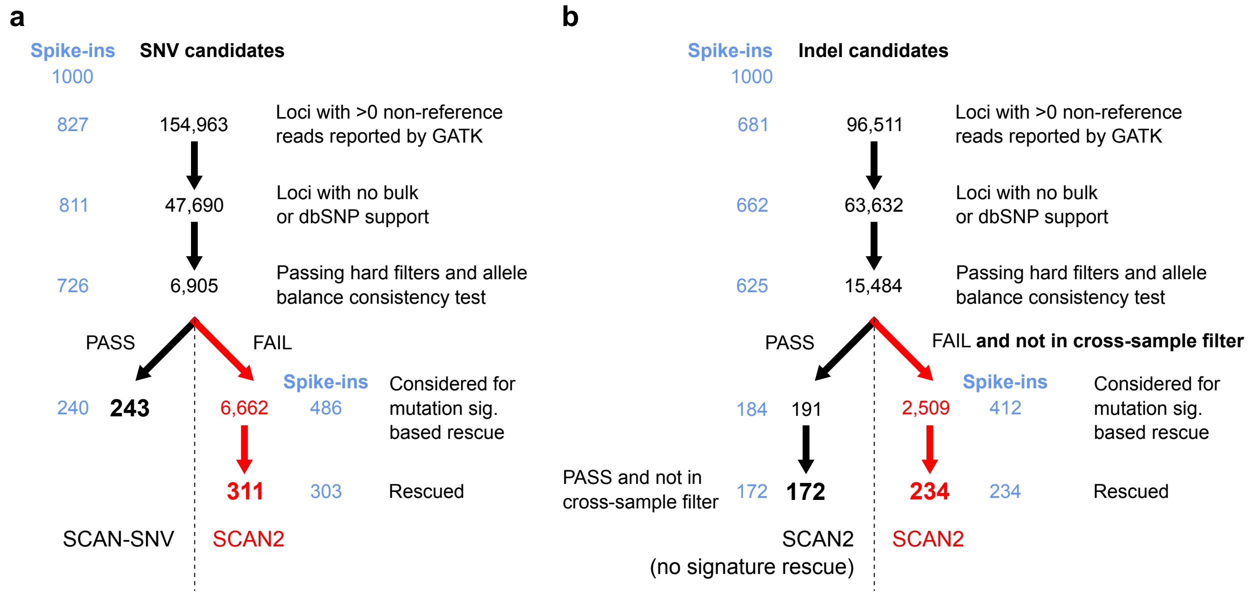
produced *N*=2 signatures corresponding to the known neuronal aging signature[6] and the universal PTA artifact signature. **c.** The SBS96 universal PTA artifact signature in more detail. **d.** Percent of SNVs in each set assigned to the artifact signature by *de novo* extraction. Values (top, *n*) indicate the total number of SNVs in each set from the 25 PTA neurons. Dotted lines: 10% and 90%. **e.** The PTA indel artifact signature in ID83 format.



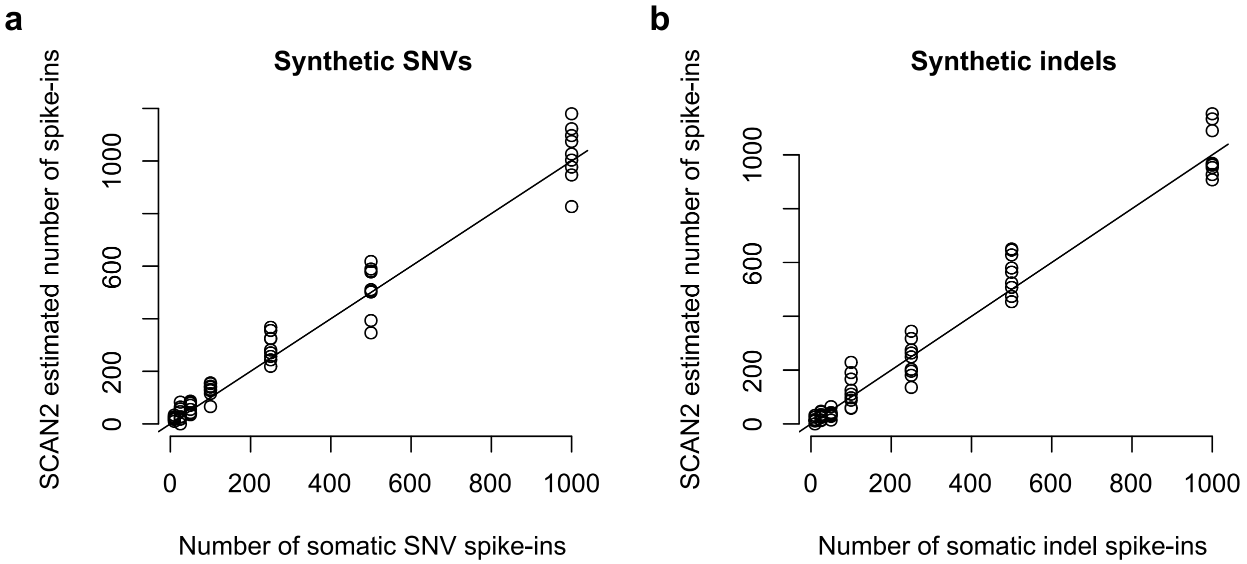
Supplementary Figure 4. Examples of the mutation signature-based rescue: weight calculation and quality score adjustment.
a. True mutation spectrum derived from high confidence calls in simulated data (synthetic X diploids). **b.** Universal PTA artifact spectrum (see Methods). **c-d.** Examples of multi-sample adjustment on two single cells (synthetic diploids) with differing artifact burdens. (*Top*) Exposure to the true and artifact mutation signatures derived by least squares fitting; cell-specific exposure to the artifact signature can be interpreted as an estimate of the artifact rate among sSNV candidates. (*Middle*) Log-scaled weights based on estimated artifact exposure, mutation type and trinucleotide context for a specific single cell. (*Bottom*) Adjustment of the FDR heuristic for sSNV candidates from one single cell. Each point represents one sSNV candidate being reconsidered by multi-sample calling. Quality scores are Phred-scaled. Detection threshold of Q=20 corresponds to a target FDR of 0.01. Solid lines, y=x.



Supplementary Figure 5. False discovery rate increases in low mutation burden contexts. MDA artifacts are more easily tolerated in cells with high mutation rates but can overwhelm somatic mutation burdens normally found in healthy human cells (0.1-1.0 sSNVs/Mb[5,6,9,10]). GATK performance: FP rate per Mb = 0.9, sensitivity = 0.8 taken from ref. 16. PTA with SCAN2 (mutation-signature and VAF-based calls) performance: FP rate per Mb = 0.0143, sensitivity = 0.458, taken from simulated SNVs in this study. The best-case MDA scenario assumes that SSD MDA SNV artifacts are completely unidentifiable and are thus treated exactly as true mutations; furthermore, it assumes SSD artifacts are the only errors committed. Our X chromosome estimate of ~584 SSD MDA SNV artifacts per genome yields ~0.22 artifacts/Mb = FP rate per Mb; sensitivity is not a necessary parameter since it cancels from the FDR calculation (because it applies equally to FPs and TPs in this scenario).

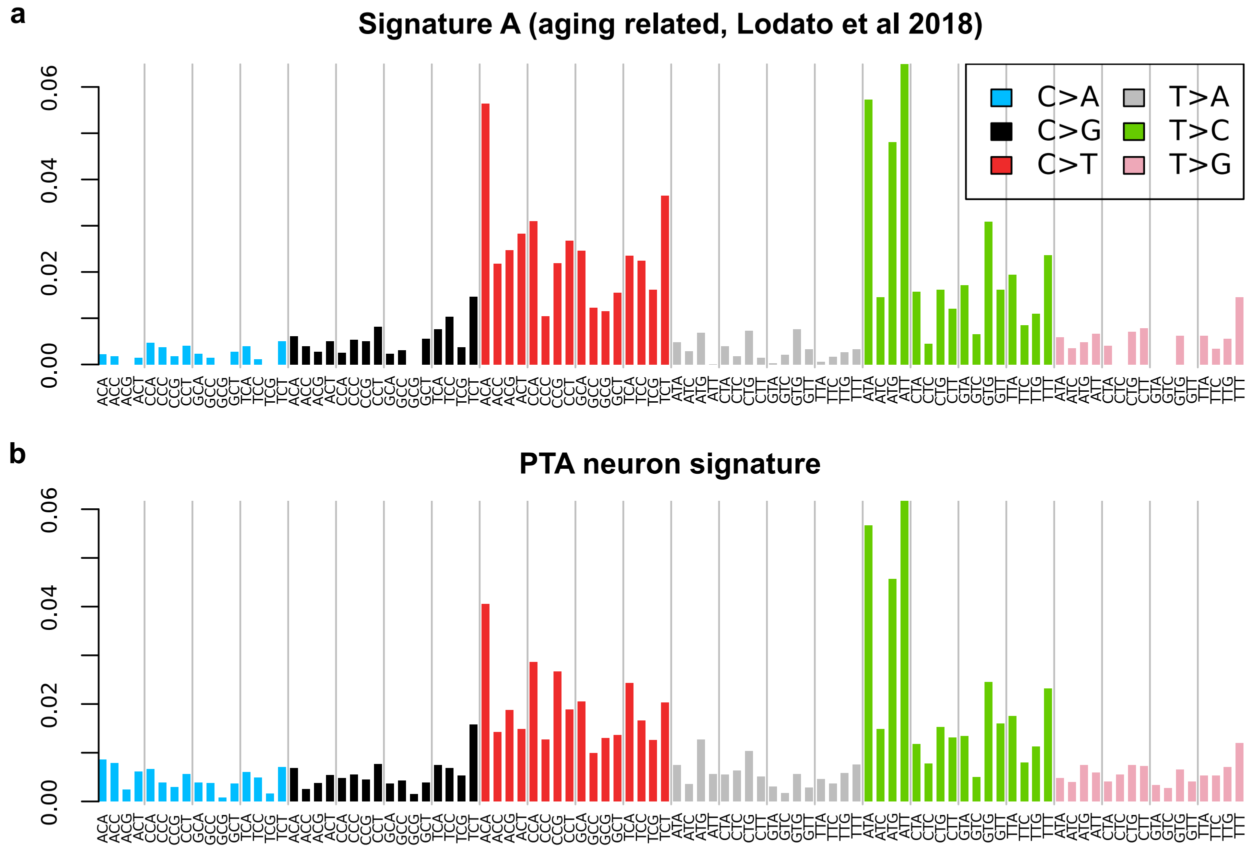

a
Spike-ins
SNV candidates
1000
827
154,963
Loci with >0 non-reference reads reported by GATK
811
47,690
Loci with no bulk or dbSNP support
726
6,905
Passing hard filters and allele balance consistency test
PASS
FAIL
240
243
6,662
Spike-ins
486
Considered for mutation sig. based rescue
311
303
Rescued
SCAN-SNV
SCAN2
b
Spike-ins
Indel candidates
1000
681
96,511
Loci with >0 non-reference reads reported by GATK
662
63,632
Loci with no bulk or dbSNP support
625
15,484
Passing hard filters and allele balance consistency test
PASS
FAIL and not in cross-sample filter
184
191
2,509
Spike-ins
412
Considered for mutation sig. based rescue
PASS and not in cross-sample filter
172
172
234
234
Rescued
SCAN2 (no signature rescue)
SCAN2

Supplementary Figure 6. Effects of SCAN2 filters on simulated data.
a. Filters for somatic SNV detection with mutation signature rescue. **b.** Filters for somatic indel detection with mutation signature rescue and cross-sample filter to remove recurrent artifacts. In both panels, values in black are the number of candidate mutations after applying the filter described on the right; light blue values are the number of simulated spike-in mutations within the set of candidate mutations. E.g., there are 827 spike-in mutations among the 154,963 SNV candidates with >0 non-reference reads reported by GATK. Red arrows and numbers show paths specific to mutation signature-based rescue. The total number of calls made by SCAN2 is the sum of the final black and red lines.



Supplementary Figure 7. SCAN2 extrapolation of total somatic mutation burden on simulated data.

By approximating somatic sensitivity by sensitivity at germline mutations, the number of called somatic mutations can be extrapolated to estimate the total number of somatic SNVs (**a**) and indels (**b**) in the cell. The method is assessed on the same synthetic diploids used for performance assessment and additional synthetic diploids with 500 and 1000 spike-ins per X chromosome.

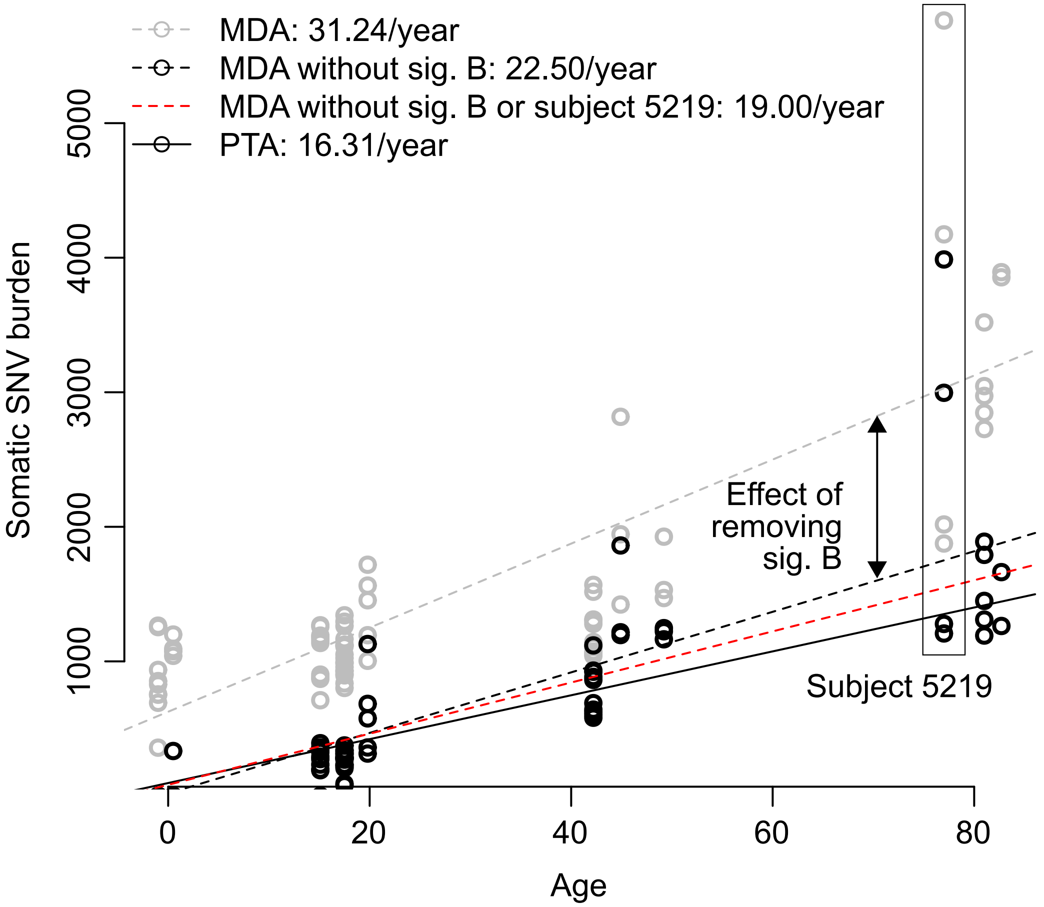


Signature A (aging related, Lodato et al 2018)

PTA neuron signature

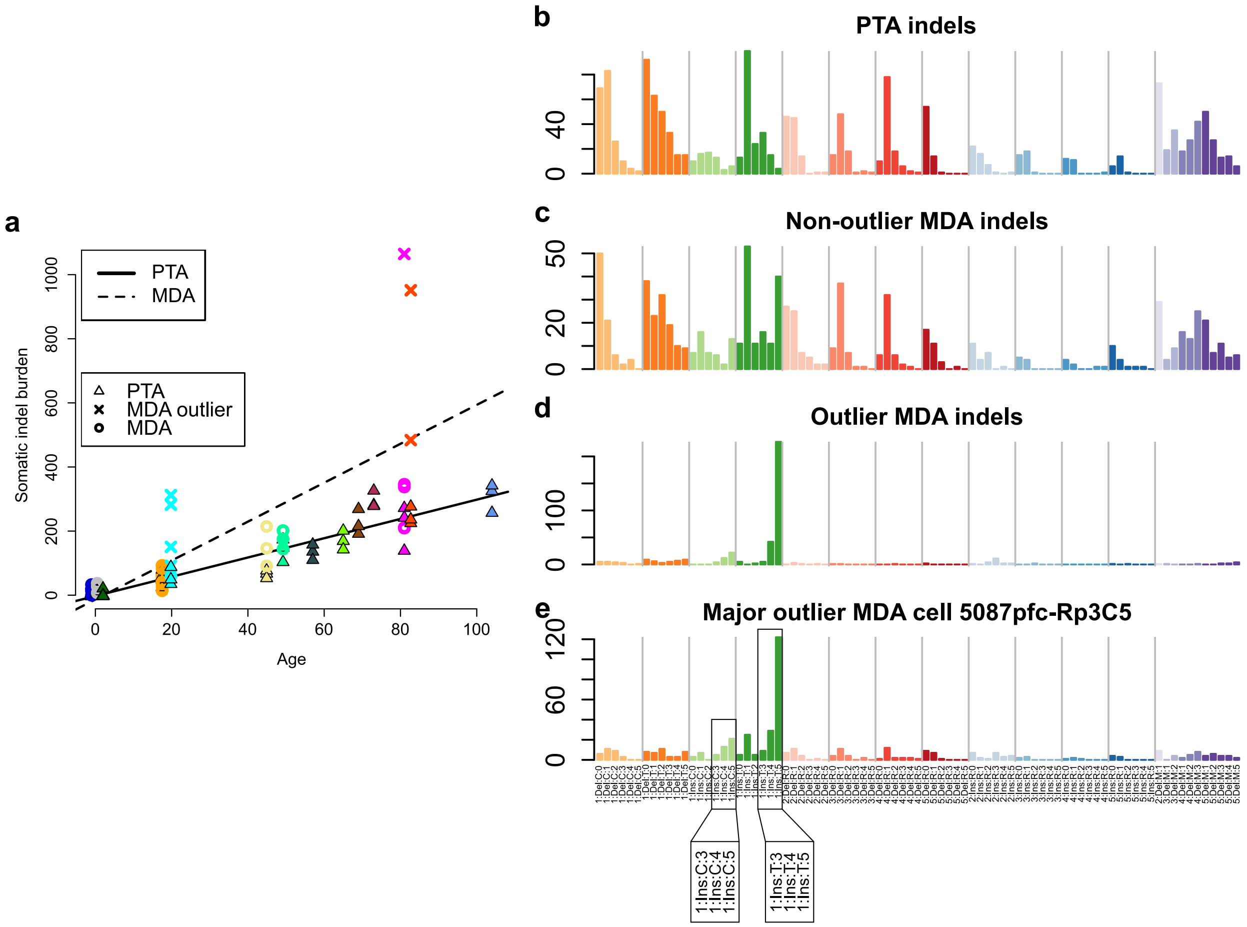
Supplementary Figure 8. PTA confirms the age-related sSNV signature in human neurons.

a. Aging-associated signature derived from MDA-amplified neurons (ref. 6)**. b.** Sole mutation signature produced by *de novo* signature extraction on PTA-amplified neurons. Only VAF-based calls were analyzed in this extraction; signature-rescued SCAN2 calls are not optimal for mutation signature analysis due to bias against mutations from signature channels with high representation in the universal PTA artifact signature. The PTA neuronal signature is highly similar to Signature A (cosine similarity=0.966), confirming the previously reported signature.

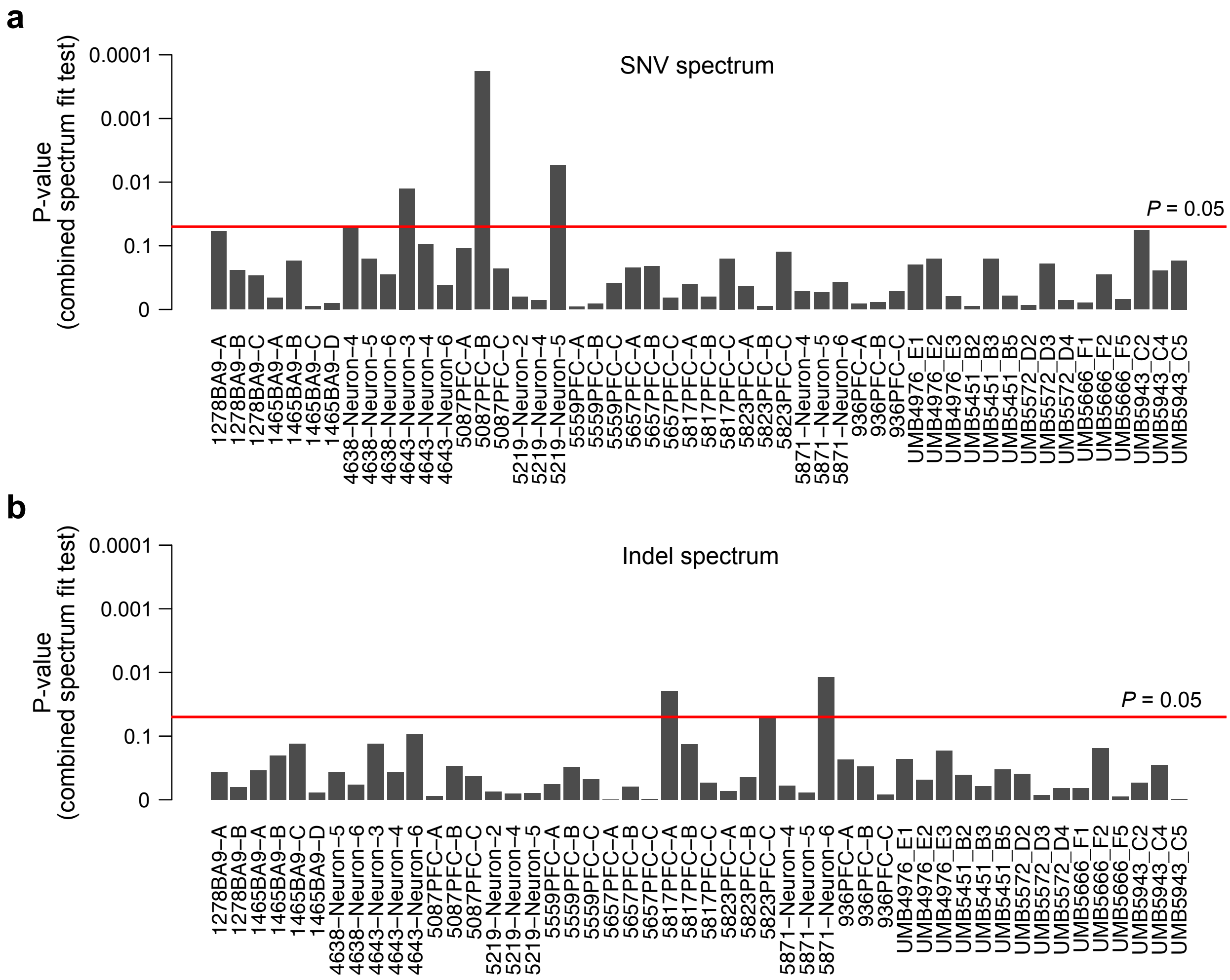


Supplementary Figure 9. Removal of Signature B from MDA neurons closely matches PTA-derived mutation rates.

Total SCAN2-called somatic SNV mutation burdens from MDA neurons before Signature B removal (grey circles) and after Signature B removal (black circles). Trend lines: MDA accumulation rate (dotted grey), MDA accumulation rate after Signature B removal (dotted black), MDA accumulation rate after Signature B and subject 5219 removal (dotted red), PTA accumulation rate (solid black).



Supplementary Figure 10. Somatic indel analysis of MDA amplified neurons.
a. Rate of somatic indel accumulation with age in MDA and PTA cells. 7 MDA-amplified neurons that are identified as outliers due to high mutation burden are represented by crosses. Linear regressions include outliers. **b-e.** ID83 signatures of PTA indels (b), MDA indels from non-outlier cells (c), indels from 7 MDA outliers (d) and indels from a single MDA outlier cell that was not included in either SNV or indel analysis (e, 5087pfc-Rp3C5). MDA neurons from subjects 4638, 4643 and 5219 were not analyzed for indels.



Supplementary Figure 11. SCAN2 batch-wide signature homogeneity test for PTA neurons in this study.

The signature homogeneity test determines whether it is appropriate to combine mutations from multiple single cells together to improve the estimation of the true mutation spectrum for signature-based calling. P-values test each single cell's SNV (**a**) or indel (**b**) spectrum against the respective batch-wide spectrum. Neurons 1278BA9-C and 4638-Neuron-4 had no indel calls and thus are not included in panel b.

Subject ID	Age	Sex	MDA	PTA
Infant				
1278	0.4	M	9	3
5817	0.6	M	4	3
5871	2.0	M	0	3
Adolescent				
4638	15.1	F	11	3
1465	17.5	M	18	4
5559	19.8	F	5	3
Adult				
4643	42.2	F	10	3
5087	44.9	M	3	3
936	49.2	F	3	3
UMB5451	57.0	F	0	3
Aged				
UMB5666	65.0	M	0	3
UMB5943	69.0	M	0	3
UMB5572	70.0	F	0	3
5219	77.0	F	4	3
5657	82	M	5	3
5823	82.7	F	3	3
UMB4976	104.0	F	0	3
			76	**52**

Supplementary Table 1: Individuals sequenced in this study. Individuals from four age groups, ranging from infants to the elderly, were analyzed in this study. MDA and PTA columns refer to the number of PFC neurons amplified by each method and sequenced to high coverage.

Subject	Sample ID	Amp.	Age	Sex	MAPD
1278	1278BA9-A	PTA	0.4	M	0.1879407
1278	1278BA9-B	PTA	0.4	M	0.1869505
1278	1278BA9-C	PTA	0.4	M	0.1860633
1278	1278_ct_p1E3	MDA	0.4	M	0.5821706
1278	1278_ct_p1E6	MDA	0.4	M	0.7670143
1278	1278_ct_p1G9	MDA	0.4	M	0.7168447
1278	1278_ct_p2B9	MDA	0.4	M	0.7078013
1278	1278_ct_p2C7	MDA	0.4	M	0.7270307
1278	1278_ct_p2E4	MDA	0.4	M	0.7440151
1278	1278_ct_p2E6	MDA	0.4	M	0.7295877
1278	1278_ct_p2F5	MDA	0.4	M	0.7097737
1278	1278_ct_p2G5	MDA	0.4	M	0.7216489
1278	1278_heart_bulk	bulk	0.4	M	0.07575756
5817	5817PFC-A	PTA	0.6	M	0.2315732
5817	5817PFC-B	PTA	0.6	M	0.2263779
5817	5817PFC-C	PTA	0.6	M	0.2142402
5817	5817_ct_p1H10	MDA	0.6	M	0.8270598
5817	5817_ct_p1H2	MDA	0.6	M	0.7535527
5817	5817_ct_p1H5	MDA	0.6	M	0.7533849
5817	5817_ct_p2H6	MDA	0.6	M	0.767587
5817	5817_liver_bulk	bulk	0.6	M	0.05570766
5871	5871-Neuron-4	PTA	2	M	0.2150414
5871	5871-Neuron-5	PTA	2	M	0.1959879
5871	5871-Neuron-6	PTA	2	M	0.1985626
5871	5871-BLK-liver	bulk	2	M	0.06179103
4638	4638-Neuron-4	PTA	15.1	F	0.1704697
4638	4638-Neuron-5	PTA	15.1	F	0.189251
4638	4638-Neuron-6	PTA	15.1	F	0.1959974
4638	4638-MDA-2	MDA	15.1	F	0.5707111
4638	4638-MDA-03	MDA	15.1	F	0.5584856
4638	4638-MDA-4	MDA	15.1	F	0.6337621
4638	4638-MDA-7	MDA	15.1	F	0.5743504
4638	4638-MDA-11	MDA	15.1	F	0.6784746
4638	4638-MDA-12	MDA	15.1	F	0.6234228
4638	4638-MDA-13	MDA	15.1	F	0.7057524
4638	4638-MDA-14	MDA	15.1	F	0.7057524
4638	4638-MDA-15	MDA	15.1	F	0.6371107
4638	4638-MDA-20	MDA	15.1	F	0.5915797
4638	4638-MDA-24	MDA	15.1	F	0.6842562
4638	4638-Bulk-Heart	bulk	15.1	F	0.07224365
1465	1465BA9-A	PTA	17.5	M	0.20723
1465	1465BA9-B	PTA	17.5	M	0.2795143
1465	1465BA9-C	PTA	17.5	M	0.2318285

1465	1465BA9-D	PTA	17.5	M	0.2719214
1465	1465-cortex_1-neuron_MDA_12	MDA	17.5	M	0.576046
1465	1465-cortex_1-neuron_MDA_18	MDA	17.5	M	0.5686175
1465	1465-cortex_1-neuron_MDA_20	MDA	17.5	M	0.548979
1465	1465-cortex_1-neuron_MDA_24	MDA	17.5	M	0.6301612
1465	1465-cortex_1-neuron_MDA_25	MDA	17.5	M	0.5743245
1465	1465-cortex_1-neuron_MDA_2_WGSb	MDA	17.5	M	0.5529697
1465	1465-cortex_1-neuron_MDA_30	MDA	17.5	M	0.6074113
1465	1465-cortex_1-neuron_MDA_39	MDA	17.5	M	0.6100418
1465	1465-cortex_1-neuron_MDA_3_WGSb	MDA	17.5	M	0.5804042
1465	1465-cortex_1-neuron_MDA_43	MDA	17.5	M	0.5449198
1465	1465-cortex_1-neuron_MDA_46	MDA	17.5	M	0.5652653
1465	1465-cortex_1-neuron_MDA_47	MDA	17.5	M	0.5699173
1465	1465-cortex_1-neuron_MDA_5	MDA	17.5	M	0.5785908
1465	1465-cortex_1-neuron_MDA_51_WGSb	MDA	17.5	M	0.6023635
1465	1465-cortex_1-neuron_MDA_6_WGSb	MDA	17.5	M	0.5610163
1465	1465-cortex_1-neuron_MDA_8	MDA	17.5	M	0.6279063
1465	1465_ct_8p2h8	MDA	17.5	M	0.5476473
1465	1465_ct_p2B11	MDA	17.5	M	0.6286406
1465	1465_ctx_p2F06	MDA	17.5	M	0.545562
1465	1465_ctx_p2g8	MDA	17.5	M	0.580386
1465	1465-heart_BulkDNA_WGSb	bulk	17.5	M	0.06634446
5559	5559PFC-A	PTA	19.8	F	0.2096737
5559	5559PFC-B	PTA	19.8	F	0.2058586
5559	5559PFC-C	PTA	19.8	F	0.1930802
5559	5559-pfc1C4	MDA	19.8	F	0.6963742
5559	5559-pfc1C7	MDA	19.8	F	0.8191077
5559	5559-pfc1E2	MDA	19.8	F	0.8608424
5559	5559-pfc1H2	MDA	19.8	F	0.8149334
5559	5559-pfc2A3	MDA	19.8	F	0.7011334
5559	5559-bulk	bulk	19.8	F	0.06573273
4643	4643-Neuron-3	PTA	42.2	F	0.1959507
4643	4643-Neuron-4	PTA	42.2	F	0.1893179
4643	4643-Neuron-6	PTA	42.2	F	0.1913095
4643	4643_MDA_1	MDA	42.2	F	0.5412151
4643	4643_MDA_2	MDA	42.2	F	0.5905864
4643	4643-MDA_23	MDA	42.2	F	0.5398771
4643	4643_MDA_24	MDA	42.2	F	0.5223155
4643	4643_MDA_26	MDA	42.2	F	0.5620214
4643	4643_MDA_3	MDA	42.2	F	0.5921091
4643	4643_MDA_31	MDA	42.2	F	0.6046078

4643	4643_MDA_32	MDA	42.2	F	0.512075
4643	4643_MDA_4	MDA	42.2	F	0.5478128
4643	4643_MDA_5	MDA	42.2	F	0.558466
4643	4643_Bulk-Liver	bulk	42.2	F	0.08035712
5087	5087PFC-A	PTA	44.9	M	0.1918068
5087	5087PFC-B	PTA	44.9	M	0.1988375
5087	5087PFC-C	PTA	44.9	M	0.1937464
5087	5087pfc-Lp1C5	MDA	44.9	M	1.105343
5087	5087pfc-Rp1G4	MDA	44.9	M	1.027076
5087	5087pfc-Rp3C5	MDA	44.9	M	1.758147
5087	5087pfc-Rp3F4	MDA	44.9	M	0.8477398
5087	5087-hrt-1b1	bulk	44.9	M	0.06029814
936	936PFC-A	PTA	49.2	F	0.1886625
936	936PFC-B	PTA	49.2	F	0.1864732
936	936PFC-C	PTA	49.2	F	0.1834495
936	936_20141001-pfc-1cp1G11_20170221-WGS	MDA	49.2	F	0.9503964
936	936_20141001-pfc-1cp1H9_20170221-WGS	MDA	49.2	F	0.8499703
936	936_20141001-pfc-1cp2F6_20170221-WGS	MDA	49.2	F	1.036471
936	936-hrt-1b1_20170221-WGS	bulk	49.2	F	0.06044224
UMB5451	UMB5451_B2*	PTA	57	F	0.221287
UMB5451	UMB5451_B3*	PTA	57	F	0.2187886
UMB5451	UMB5451_B5*	PTA	57	F	0.1962061
UMB5451	5451-190613-ctxBA4*	bulk	57	F	0.06974126
UMB5666	UMB5666_F1*	PTA	65	M	0.1969526
UMB5666	UMB5666_F2*	PTA	65	M	0.1918533
UMB5666	UMB5666_F5*	PTA	65	M	0.1926559
UMB5666	UMB5666_bulk*	bulk	65	M	0.06421471
UMB5943	UMB5943_C2*	PTA	69	M	0.1903532
UMB5943	UMB5943_C4*	PTA	69	M	0.1900218
UMB5943	UMB5943_C5*	PTA	69	M	0.1942961
UMB5943	5943-190613-ctxBA4*	bulk	69	M	0.06822946
UMB5572	UMB5572_D2*	PTA	70	F	0.198134
UMB5572	UMB5572_D3*	PTA	70	F	0.1926567
UMB5572	UMB5572_D4*	PTA	70	F	0.190076
UMB5572	UMB5572_bulk*	bulk	70	F	0.06316964
5219	5219-Neuron-2	PTA	77	F	0.1925253
5219	5219-Neuron-4	PTA	77	F	0.1910372
5219	5219-Neuron-5	PTA	77	F	0.1863487
5219	5219_ct_p1G1	MDA	77	F	0.9505295
5219	5219_ct_p1G7	MDA	77	F	0.7595051
5219	5219_ct_p2A12	MDA	77	F	1.074946
5219	5219_ct_p2C3	MDA	77	F	1.027055

5219	5219_cb_bulk	bulk	77	F	0.06258775
5657	5657PFC-A	PTA	82	M	0.1912027
5657	5657PFC-B	PTA	82	M	0.1868522
5657	5657PFC-C	PTA	82	M	0.1845906
5657	5657-pfc1D2	MDA	82	M	1.076119
5657	5657-pfc1E11	MDA	82	M	0.7708719
5657	5657-pfc2A6	MDA	82	M	0.740058
5657	5657-pfc2F1	MDA	82	M	0.7283375
5657	5657-pfc2G9	MDA	82	M	0.7480412
5657	5657-bulk	bulk	82	M	0.06028281
5823	5823PFC-A	PTA	82.7	F	0.1940577
5823	5823PFC-B	PTA	82.7	F	0.2153974
5823	5823PFC-C	PTA	82.7	F	0.189505
5823	5823_20160824-pfc-1cp2E1_20170221-WGS	MDA	82.7	F	1.143267
5823	5823_20160824-pfc-1cp2G5_20170221-WGS	MDA	82.7	F	1.062387
5823	5823-tempmusc-1b1_20170221-WGS	bulk	82.7	F	0.05498299
UMB4976	UMB4976_E1*	PTA	104	F	0.2016462
UMB4976	UMB4976_E2*	PTA	104	F	0.1910937
UMB4976	UMB4976_E3*	PTA	104	F	0.195309
UMB4976	4976-190613-cer*	bulk	104	F	0.06182438

Supplementary Table 2: Samples analyzed in this study. List of all samples used in this study. For single cell samples, the method of genome amplification is listed (MDA or PTA); samples with amplification “none” are bulk controls. Callable bp indicates the number of base pairs in the human genome which passed basic depth criteria for analysis (>5 in the single cell, >10 in the matched bulk). Additional PTA cells and bulks marked with * were generated in ref. 20.

Conda package name	Version	Conda extra version
_libgcc_mutex	0.1	main
_r-mutex	1.0.0	anacondar_1
aioeasywebdav	2.2.0	py36_0
aiohttp	3.4.4	py36h470a237_0
anaconda-client	1.7.1	py_0
appdirs	1.4.3	py_1
asn1crypto	0.24.0	py36_1003
async-timeout	3.0.1	py_1000
attrs	18.2.0	py_0
bcrypt	3.1.4	py36h470a237_0
beautifulsoup4	4.6.3	py36_1000
bedtools	2.27.1	he941832_2
bioconductor-annotationdbi	1.42.1	r351_0
bioconductor-annotationfilter	1.4.0	r351_0
bioconductor-biobase	2.40.0	r351ha44fe06_1
bioconductor-biocgenerics	0.26.0	r351_0
bioconductor-biocparallel	1.14.2	r351h26a2512_0
bioconductor-biomart	2.36.1	r351_0
bioconductor-biostrings	2.48.0	r351h470a237_0
bioconductor-biovizbase	1.28.2	r351h470a237_0
bioconductor-bsgenome	1.48.0	r351_0
bioconductor-bsgenome.hsapiens.ucsc.hg19	1.4.0	r351_4
bioconductor-delayedarray	0.6.6	r351_0
bioconductor-ensembldb	2.4.1	r351_0
bioconductor-genomeinfodb	1.16.0	r351_0
bioconductor-genomeinfodbdata	1.1.0	r351_0
bioconductor-genomicalignments	1.16.0	r351h470a237_0
bioconductor-genomicfeatures	1.32.3	r351_0
bioconductor-genomicranges	1.32.7	r351h470a237_0
bioconductor-gviz	1.24.0	r351_0
bioconductor-iranges	2.14.12	r351h470a237_0
bioconductor-protgenerics	1.12.0	r351_0
bioconductor-rsamtools	1.32.3	r351hfc679d8_0
bioconductor-rtracklayer	1.40.6	r351h470a237_0
bioconductor-s4vectors	0.18.3	r351h470a237_0
bioconductor-summarizedexperiment	1.10.1	r351_0
bioconductor-variantannotation	1.26.1	r351h470a237_0
bioconductor-xvector	0.20.0	r351h470a237_0

bioconductor-zlibbioc	1.26.0	r351h470a237_0
blas	1	mkl
boost	1.66.0	py36_1
boost-cpp	1.66.0	1
boto3	1.7.84	py_0
botocore	1.10.84	py_0
bwidget	1.9.11	1
bzip2	1.0.6	h470a237_2
ca-certificates	2018.11.29	ha4d7672_0
cachetools	2.1.0	py_0
cairo	1.14.12	h276e583_5
certifi	2018.11.29	py36_1000
cffi	1.11.5	py36h5e8e0c9_1
chardet	3.0.4	py36_1003
clyent	1.2.2	py_1
conda	4.5.12	py36_1000
conda-build	3.17.5	py36_0
conda-env	2.6.0	1
configargparse	0.13.0	py_1
cryptography	2.3.1	py36hdffb7b8_0
cryptography-vectors	2.3.1	py36_1000
curl	7.63.0	h74213dd_0
datrie	0.7.1	py36h7b6447c_1
decorator	4.3.0	py_0
docutils	0.14	py36_1001
dropbox	7.3.1	py36_0
eagle-phase	2.3.5	0
expat	2.2.5	hfc679d8_2
filechunkio	1.6	py36_0
filelock	3.0.10	py_0
fontconfig	2.13.1	h65d0f4c_0
freetype	2.9.1	h6debe1e_4
ftputil	3.2	py36_0
gatk	3.8	py36_0
gettext	0.19.8.1	h5e8e0c9_1
gitdb2	2.0.5	py_0
gitpython	2.1.11	py_0
glib	2.56.2	h464dc38_1
glob2	0.6	py_0
google-auth	1.2.1	py_0

google-auth-httplib2	0.0.3	py_2
google-cloud-core	0.24.1	py36_0
google-cloud-storage	1.1.1	py36_0
google-resumable-media	0.0.2	py36_0
googleapis-common-protos	1.5.5	py_0
graphite2	1.3.12	hfc679d8_1
graphviz	2.38.0	h08bfae6_9
gsl	2.2.1	h0c605f7_3
harfbuzz	1.9.0	h04dbb29_1
htslib	1.9	h47928c2_5
httplib2	0.12.0	py36_1000
icu	58.2	hfc679d8_0
idna	2.8	py36_1000
idna_ssl	1.0.0	0
intel-openmp	2019.1	144
ipython_genutils	0.2.0	py_1
jinja2	2.1	py_1
jmespath	0.9.3	py_1
jpeg	9c	h470a237_1
jsonschema	3.0.0a3	py36_1000
jupyter_core	4.4.0	py_0
krb5	1.16.2	hbb41f41_0
libarchive	3.3.3	h823be47_0
libcurl	7.63.0	hbdb9355_0
libdeflate	1	h470a237_0
libedit	3.1.20170329	h6b74fdf_2
libffi	3.2.1	hd88cf55_4
libgcc-ng	8.2.0	hdf63c60_1
libgfortran	3.0.0	1
libgfortran-ng	7.2.0	hdf63c60_3
libiconv	1.15	h470a237_3
liblief	0.9.0	h7725739_1
libpng	1.6.36	ha92aebf_0
libprotobuf	3.6.1	hd28b015_0
libssh2	1.8.0	h5b517e9_3
libstdcxx-ng	8.2.0	hdf63c60_1
libtiff	4.0.10	he6b73bb_1
libtool	2.4.6	h470a237_2
libuuid	2.32.1	h470a237_2
libxcb	1.13	h470a237_2

libxml2	2.9.8	h422b904_5
make	4.2.1	h470a237_1002
markupsafe	1.1.0	py36h470a237_0
mkl	2019.1	144
mkl_fft	1.0.10	py36_0
mkl_random	1.0.2	py36_0
multidict	4.5.1	py36h470a237_0
nbformat	4.4.0	py_1
ncurses	6.1	hf484d3e_0
networkx	2.2	py_1
numpy	1.15.4	py36h7e9f1db_0
numpy-base	1.15.4	py36hde5b4d6_0
openblas	0.3.4	ha44fe06_0
openjdk	8.0.192	h470a237_2
openssl	1.0.2p	h470a237_1
pandas	0.23.4	py36hf8a1672_0
pango	1.40.14	he752989_2
paramiko	2.4.2	py36_1000
patchelf	0.9	hfc679d8_2
pcre	8.41	hfc679d8_3
picard	2.18.23	0
pip	18.1	py36_1000
pixman	0.34.0	h470a237_3
pkginfo	1.4.2	py_1
prettytable	0.7.2	py_2
protobuf	3.6.1	py36hfc679d8_1
psutil	5.4.8	py36h470a237_0
pthread-stubs	0.4	h470a237_1
py-lief	0.9.0	py36h7725739_1
pyasn1	0.4.4	py_1
pyasn1-modules	0.0.5	py36_0
pycosat	0.6.3	py36h470a237_1
pycparser	2.19	py_0
pygraphviz	1.3.1	py36_0
pynacl	1.3.0	py36h470a237_0
pyopenssl	18.0.0	py36_1000
pyrsistent	0.14.7	py36h470a237_0
pysftp	0.2.9	py36_0
pysocks	1.6.8	py36_1002
python	3.6.6	h5001a0f_3

python-dateutil	2.7.5	py_0
python-irodsclient	0.7.0	py_0
python-libarchive-c	2.8	py36_1004
pytz	2018.7	py_0
pyyaml	3.13	py36h470a237_1
r-acepack	1.4.1	r35h9bbef5b_1004
r-assertthat	0.2.0	r351h6115d3f_1001
r-backports	1.1.5	r35hcdcec82_0
r-base	3.5.1	h391c2eb_4
r-base64enc	0.1_3	r35hcdcec82_1003
r-bh	1.66.0_1	r351_2001
r-bindr	0.1.1	r351h6115d3f_1001
r-bindrcpp	0.2.2	r351h9d2a408_1
r-bit	1.1_12	r351h470a237_2
r-bit64	0.9_7	r351hc070d10_0
r-bitops	1.0_6	r351hc070d10_2
r-blob	1.1.1	r351_1001
r-callr	3.4.2	r35h6115d3f_0
r-checkmate	2.0.0	r35hcdcec82_0
r-cli	1.0.1	r351h6115d3f_1000
r-clipr	0.7.0	r35h6115d3f_0
r-clisymbols	1.2.0	r35h6115d3f_1002
r-cluster	2.1.0	r35h9bbef5b_2
r-colorspace	1.4_1	r35hcdcec82_1
r-crayon	1.3.4	r351h6115d3f_1001
r-curl	4.3	r35hcdcec82_0
r-data.table	1.12.8	r35hcdcec82_0
r-dbi	1.0.0	r351h6115d3f_1001
r-dbplyr	1.2.2	r351h6115d3f_1001
r-desc	1.2.0	r35h6115d3f_1002
r-devtools	2.0.2	r351h6115d3f_0
r-dichromat	2.0_0	r35_2001
r-digest	0.6.18	r351hc070d10_0
r-dplyr	0.7.8	r351h9d2a408_0
r-evaluate	0.14	r35h6115d3f_1
r-fansi	0.3.0	r351hc070d10_0
r-fastghquad	1	r351h29659fb_0
r-foreign	0.8_76	r35hcdcec82_0
r-formatr	1.5	r351h6115d3f_1001
r-formula	1.2_3	r35h6115d3f_1002

r-fs	1.3.2	r35h0357c0b_0
r-futile.logger	1.4.3	r351h6115d3f_1001
r-futile.options	1.0.1	r351h6115d3f_1000
r-ggplot2	3.3.0	r35h6115d3f_0
r-gh	1.1.0	r35h6115d3f_0
r-git2r	0.24.0	r351h47c54a8_1
r-glue	1.3.0	r351h470a237_2
r-gridextra	2.3	r35h6115d3f_1002
r-gtable	0.3.0	r35h6115d3f_2
r-highr	0.8	r35h6115d3f_1
r-hmisc	4.4_0	r35h9bbef5b_0
r-hms	0.4.2	r351h6115d3f_1000
r-htmltable	1.13.3	r35h6115d3f_0
r-htmltools	0.4.0	r35h0357c0b_0
r-htmlwidgets	1.5.1	r35h6115d3f_0
r-httr	1.4.1	r35h6115d3f_1
r-ini	0.3.1	r35h6115d3f_1002
r-isoband	0.2.1	r35h0357c0b_0
r-jsonlite	1.6.1	r35hcdcec82_0
r-knitr	1.28	r35h6115d3f_0
r-labeling	0.3	r35h6115d3f_1002
r-lambda.r	1.2.3	r351h6115d3f_1000
r-lattice	0.20_38	r351hc070d10_0
r-latticeextra	0.6_28	r35h6115d3f_1002
r-lazyeval	0.2.2	r35hcdcec82_1
r-magrittr	1.5	r351h6115d3f_1001
r-markdown	1.1	r35hcdcec82_0
r-mass	7.3_51.5	r35hcdcec82_0
r-matrix	1.2_15	r351hc070d10_0
r-matrixstats	0.54.0	r351hc070d10_0
r-memoise	1.1.0	r351h6115d3f_1001
r-mgcv	1.8_31	r35hcdcec82_0
r-mime	0.9	r35hcdcec82_0
r-munsell	0.5.0	r35h6115d3f_1002
r-nlme	3.1_147	r35h9bbef5b_0
r-nnet	7.3_13	r35hcdcec82_0
r-openssl	1.1	r351hff1dc39_1001
r-pillar	1.3.0	r351h6115d3f_1000
r-pkgbuild	1.0.6	r35h6115d3f_0
r-pkgconfig	2.0.2	r351h6115d3f_1001

r-pkgload	1.0.2	r35h0357c0b_1001
r-plogr	0.2.0	r351h6115d3f_1001
r-praise	1.0.0	r35h6115d3f_1003
r-prettyunits	1.0.2	r351h6115d3f_1001
r-processx	3.4.2	r35hcdcec82_0
r-progress	1.2.2	r35h6115d3f_1
r-ps	1.3.2	r35hcdcec82_0
r-purrr	0.2.5	r351hc070d10_2
r-r6	2.2.2	r351h6115d3f_1001
r-rappdirs	0.3.1	r35hcdcec82_1003
r-rcmdcheck	1.3.2	r351h6115d3f_1000
r-rcolorbrewer	1.1_2	r35h6115d3f_1002
r-rcpp	1.0.0	r351h9d2a408_0
r-rcurl	1.95_4.11	r351hc070d10_3
r-remotes	2.1.1	r35h6115d3f_0
r-reticulate	1.14	r35h0357c0b_0
r-rlang	0.3.0.1	r351h470a237_0
r-rpart	4.1_15	r35hcdcec82_1
r-rprojroot	1.3_2	r35h6115d3f_1002
r-rsqlite	2.1.1	r351h9d2a408_0
r-rstudioapi	0.11	r35h6115d3f_0
r-scales	1.0.0	r35h0357c0b_1002
r-scansnv	0.1	r351hf484d3e_0
r-sessioninfo	1.1.1	r35h6115d3f_1001
r-snow	0.4_3	r351h6115d3f_1000
r-stringi	1.4.3	r35h0357c0b_2
r-stringr	1.4.0	r35h6115d3f_1
r-survival	3.1_12	r35hcdcec82_0
r-testthat	2.2.1	r35h0357c0b_0
r-tibble	1.4.2	r351hc070d10_2
r-tidyselect	0.2.5	r351h9d2a408_0
r-usethis	1.5.1	r35h6115d3f_1
r-utf8	1.1.4	r351hc070d10_0
r-viridis	0.5.1	r35h6115d3f_1003
r-viridislite	0.3.0	r35h6115d3f_1002
r-whisker	0.4	r35h6115d3f_0
r-withr	2.1.2	r35h6115d3f_1001
r-xfun	0.13	r35h6115d3f_0
r-xml	3.98_1.16	r351hc070d10_0
r-xopen	1.0.0	r35h6115d3f_1002

r-yaml	2.2.1	r35hcdcec82_0
ratelimiter	1.2.0	py36_1000
readline	7	h7b6447c_5
requests	2.13.0	py36_0
rpy2	2.9.4	py36r351h941a26a_1
rsa	3.1.4	py36_0
ruamel_yaml	0.15.71	py36h470a237_0
s3transfer	0.1.13	py36_1001
samtools	1.9	h8ee4bcc_1
scansnv	1	0
setuptools	40.6.3	py36_0
shapeit	2.r837	0
shyaml	0.6.1	py_0
six	1.12.0	py36_1000
smmap2	2.0.5	py_0
snakemake	5.3.1	0
snakemake-minimal	5.3.1	py_0
sqlite	3.25.3	h7b6447c_0
tk	8.6.8	hbc83047_0
tktable	2.1	h14c3975_0
tqdm	4.28.1	py_0
traitlets	4.3.2	py36_1000
tzlocal	1.5.1	py_0
urllib3	1.12	py36_0
wget	1.19.5	h1ad7b7a_0
wheel	0.32.3	py36_0
wrapt	1.10.11	py36h470a237_1
xmlrunner	1.7.7	py_0
xorg-kbproto	1.0.7	h470a237_2
xorg-libice	1.0.9	h470a237_4
xorg-libsm	1.2.3	h8c8a85c_0
xorg-libx11	1.6.6	h470a237_0
xorg-libxau	1.0.8	h470a237_6
xorg-libxdmcp	1.1.2	h470a237_7
xorg-libxext	1.3.3	h470a237_4
xorg-libxpm	3.5.12	h470a237_2
xorg-libxrender	0.9.10	h470a237_2
xorg-libxt	1.1.5	h470a237_2
xorg-renderproto	0.11.1	h470a237_2
xorg-xextproto	7.3.0	h470a237_2

xorg-xproto	7.0.31	h470a237_7
xz	5.2.4	h14c3975_4
yaml	0.1.7	had09818_2
yarl	1.3.0	py36h470a237_0
zlib	1.2.11	ha838bed_2
cycler	0.10.0	
drmaa	0.7.9	
kiwisolver	1.1.0	
matplotlib	3.2.1	
patsy	0.5.1	
pyparsing	2.4.6	
scipy	1.4.1	
sigprofilermatrixgenerator	1.1.9	
sigprofilerplotting	1.1.0	
statsmodels	0.11.1	

Supplementary Table 3: Software packages used in this study. List of all packages present in the conda environment used in this study. This table is also available as a conda-formatted environment file