

# 1 **Ultraspecific somatic SNV and indel detection in single** 2 **neurons using primary template-directed amplification**

3  
4 Lovelace J. Luquette<sup>1,\*</sup>, Michael B. Miller<sup>2,3,4,\*</sup>, Zinan Zhou<sup>2,\*</sup>, Craig L. Bohrson<sup>1</sup>, Alon Galor<sup>1</sup>,  
5 Michael A. Lodato<sup>5</sup>, Charles Gawad<sup>6,7</sup>, Jay West<sup>8</sup>, Christopher A. Walsh<sup>2,3,9,§</sup> and Peter J.  
6 Park<sup>1,10,§</sup>

7  
8 <sup>1</sup> Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

9 <sup>2</sup> Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA.

10 <sup>3</sup> Manton Center for Orphan Disease, Boston Children's Hospital, Boston, MA, USA;  
11 Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA; and Broad  
12 Institute of MIT and Harvard, Cambridge, MA, USA.

13 <sup>4</sup> Division of Neuropathology, Department of Pathology, Brigham and Women's Hospital,  
14 Harvard Medical School, Boston, MA, USA.

15 <sup>5</sup> Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical  
16 School, Worcester, MA, USA.

17 <sup>6</sup> Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA.

18 <sup>7</sup> Chan Zuckerberg Biohub, San Francisco, CA, USA.

19 <sup>8</sup> BioSkryb, Durham, NC, USA.

20 <sup>9</sup> Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA.

21 <sup>10</sup> Ludwig Center at Harvard, Boston, MA, USA.

22  
23  
24 \* These authors contributed equally to this work.

25 § These authors jointly supervised this work.

## 26 27 28 **Abstract**

29 Primary template-directed amplification (PTA) is an improved amplification technique for  
30 single-cell DNA sequencing. We generated whole-genome analysis of 76 single neurons and  
31 developed SCAN2, a computational method to accurately identify both clonal and non-clonal  
32 somatic (i.e., limited to a single neuron) single nucleotide variants (SNVs) and small insertions  
33 and deletions (indels) using PTA data. Our analysis confirms an increase in non-clonal somatic  
34 mutation in single neurons with age, but revises estimates for the rate of this accumulation to  
35 be 15 SNVs per year. We also identify artifacts in other amplification methods. Most  
36 importantly, we show that somatic indels also increase by at least 2 indels per year per neuron  
37 and that indels may have a larger impact on gene function than somatic SNVs in human  
38 neurons.

## 42 Introduction

43 Although somatic mutation has been studied extensively in cancer, investigation into the  
44 abundance, patterns, and effects of somatic mosaicism in non-neoplastic tissues has only  
45 recently begun<sup>1-6</sup>. Unlike tumor tissue in which somatic mutations of interest are shared by  
46 large clones, the majority of somatic mutations in normal tissues are typically shared by  
47 relatively few cells and are therefore difficult to detect. Recent studies have circumvented the  
48 technical difficulty of detecting rare somatic mutations by strategies including ultradeep  
49 sequencing of very small tissue samples<sup>3,7</sup>, exploiting naturally occurring genetically  
50 homogenous clones<sup>8</sup>, or clonal expansion of cells *in vitro*<sup>5,9,10</sup>.

51  
52 Another strategy for detecting somatic mosaic mutations is to directly sequence DNA from a  
53 single cell. Single cell DNA sequencing (scDNA-seq) is capable of detecting the rarest somatic  
54 mutations (i.e., mutations private to a single cell) and can also provide information about cell  
55 lineage through shared somatic mutations<sup>2,11</sup>. This strategy is especially useful for examining  
56 somatic mutations in post-mitotic cells such as neurons, in which their presence is limited to  
57 single cells. A major bottleneck, however, has been the difficulty of amplifying the genome of a  
58 single cell accurately and evenly so that it can be sequenced by a high-throughput sequencer.  
59 For example, multiple displacement amplification (MDA)<sup>12</sup>, a popular amplification method for  
60 detecting point mutations, produces non-uniformity across the genome<sup>13</sup> and often amplifies  
61 homologous alleles of diploid cells at different rates, leading to allelic imbalance<sup>14</sup>. These  
62 amplification artifacts pose substantial difficulties for identifying mutations from short-read  
63 sequencing data—especially mutations that are non-clonal and thus cannot be confirmed by  
64 sequencing multiple single cells. We previously used read-level phasing to filter artifacts in MDA  
65 samples and discovered an age-associated increase in somatic mutations in human neurons<sup>6</sup>,  
66 but were limited to analyzing mutations within a few hundred base pairs of germline SNPs  
67 (~15% of the genome). A newly developed single-cell amplification method called primary  
68 template-directed amplification (PTA) aims to reduce these artifacts by dampening the  
69 exponential nature of isothermal MDA<sup>15</sup>.

70  
71 Here we compare single neurons amplified by both the MDA and PTA protocols from the  
72 prefrontal cortices of the same individuals and find that PTA substantially improves upon MDA.  
73 Nevertheless, conventional somatic SNV analysis (based on Genome Analysis Toolkit (GATK)  
74 best practices) of PTA data yields 0.9 false positives (FPs) per megabase, exceeding the  
75 mutation rate in some non-neoplastic cells by an order of magnitude<sup>10</sup>. We therefore  
76 developed SCAN2 (Single Cell Analysis 2), a small mutation genotyper based on the SCAN-SNV<sup>14</sup>  
77 model of allelic imbalance. SCAN2 detects non-clonal somatic SNVs and indels in scDNA-seq  
78 data with 60-fold fewer FPs per megabase than conventional calling and >5-fold fewer FPs than  
79 single-cell SNV genotypers. Somatic SNV detection in SCAN2 is greatly improved by a novel  
80 multi-sample approach that distinguishes mutations from artifacts based on 96-dimensional  
81 mutation signatures<sup>16</sup>; somatic indel calling is enabled by using multiple single cells to identify  
82 and remove sites with unusually high indel recurrences. SCAN2 confirms a previously reported  
83 signature of single nucleotide MDA artifacts<sup>17</sup> and revises the rate of somatic SNV (sSNV)  
84 accumulation in aging neurons from the human prefrontal cortex<sup>6</sup>. Most notably, SCAN2  
85 provides the first characterization of somatic indels in human neurons, revealing the yearly rate

86 of indel accumulation and a bias toward genic regions. Two of four known clock-like indel  
87 signatures appear to be active in neurons; additionally, we find that aging-related neuronal  
88 indels are primarily enriched for indel signature 4 from the COSMIC catalog, a signature  
89 characterized by short deletions of 2-4 bp and with no known aetiology.

90

### 91 **PTA improves amplification quality and reduces artifact burden**

92 The genomes of 25 single neurons from the prefrontal cortex (PFC) of eight neurotypical  
93 individuals were amplified by PTA and sequenced to 30-60X (**Fig. 1a, Supplementary Fig. 1a,**  
94 **Supplementary Table 1**). Compared to MDA-amplified single neuron WGS data from the same  
95 individuals<sup>6</sup>, PTA-amplified neurons showed several favorable characteristics, including  
96 substantial reduction in coverage variability across the genome (as measured by median  
97 absolute pairwise deviation (MAPD) and visual inspection of copy number profiles) and allelic  
98 imbalance, despite being sequenced to lower depth (**Fig. 1b-d**). Allelic balance measures the  
99 evenness of amplification between homologous alleles in a diploid cell; values near 0.5 indicate  
100 successful amplification of both alleles while values near 0 or 1 indicate loss of one allele. On  
101 average, only 37% of MDA-amplified genomes exhibited balance levels in the range of 0.3-0.7  
102 compared to 68% of PTA genomes. We also found the rate of amplification failure among our  
103 PTA reactions to be low: only a single PTA neuron showed evidence of amplification failure in  
104 the form of near-complete loss of several haplotypes (**Supplementary Fig. 2**). However, we  
105 cannot rule out the possibility that *bona fide* mutations are the sources of these copy losses,  
106 meaning that none of the 25 PTA reactions failed. If these were indeed true mutations, then  
107 they are the only large-scale (>5 Mb, see Methods) copy number changes we detected in these  
108 neurons, which is unexpected given reports of pervasive copy number alterations in human  
109 neurons, especially from young individuals<sup>19,20</sup>.

110

111 Comparison of the numbers of somatic mutation calls between MDA and PTA amplified  
112 neurons from the same individual suggested specific types of artifacts introduced by MDA. In  
113 the absence of artifacts, the number of somatic calls should be similar in MDA and PTA from  
114 the same individual after correction for sensitivity, while a consistent excess of calls specific to  
115 one amplification method would indicate the presence of additional artifacts and allow  
116 estimation of the artifact rate. To measure the rates of high variant allele fraction (VAF)  
117 artifacts, we analyzed male X chromosomes since mutation detection in hemizygous regions is  
118 considerably less difficult than in diploid regions (Methods). MDA neurons displayed a median  
119 excess of 15.9 somatic SNVs and 3.7 somatic indels per haploid X chromosome, indicating that  
120 one should expect about 584 SNV and 136 indel high VAF artifacts per genome (**Fig. 1e-f,**  
121 **Supplementary Fig. 1b-c**). Notably, these MDA artifacts frequently occur with variant allele  
122 fractions (VAFs) of 100%, which is compatible with a previously proposed artifact model<sup>21</sup>  
123 involving failure to amplify either the Watson or Crick strand of the initial DNA molecule.  
124 Artifacts caused by such single-stranded dropout do not leave the telltale signs of amplification  
125 artifacts (i.e., discordantly phased reads<sup>21</sup> or improper VAFs<sup>14</sup>) and are often indistinguishable  
126 from true mutations.

127

128

## 129 **High specificity is critical for somatic mutation detection in healthy cells**

130 The importance of single-stranded dropout MDA artifacts depends on how many mutations of  
131 interest exist in the cells being analyzed. For example, since human cells contain 3-4 million  
132 germline SNVs (>1000 SNVs/Mb), several hundred artifacts would have little effect on germline  
133 SNV discovery. Indeed, in the context of germline SNV detection, we estimate a false discovery  
134 rate (FDR) of <0.1% regardless of the amplification or analysis method (**Fig. 2a**). However,  
135 estimated FDR rates for MDA and conventional analysis of PTA are unacceptable when the  
136 mutations of interest are rare, as in somatic SNV detection in healthy single cells (0.1-1.0  
137 sSNVs/Mb<sup>5,6,9,10</sup>). For MDA, we estimate a best-case scenario by assuming that the only FP  
138 errors are caused by single-stranded artifacts (see Methods). Under this assumption, we expect  
139 MDA FDRs of at least 17% (for cells with 1.0 sSNVs/Mb) to 68% (for cells with 0.1 sSNVs/MB);  
140 but in practice, higher MDA FDRs would be expected due to additional FPs from non-single-  
141 stranded artifacts. Although PTA produces fewer artifacts than MDA, single-cell-aware  
142 genotypers are critical for accurate sSNV calling in low mutation burden contexts: the  
143 conventional GATK best practices pipeline (with additional filtering) was recently estimated to  
144 produce 0.9 false positives (FPs) per megabase with ~80% sensitivity in PTA amplified cells<sup>15</sup>,  
145 corresponding to FDRs of 47% (1.0 sSNVs/MB) to 90% (0.1 sSNVs/MB) for typical healthy cells.  
146 In summary, both the optimistic MDA scenario and analysis of PTA by conventional genotypers  
147 are likely to produce unacceptable FDR levels in cells with low mutation burden. We therefore  
148 developed SCAN2, which achieves FDR < ~15% even for cells with very low mutation burden  
149 (0.1 sSNVs/Mb).

150

## 151 **SCAN2 accurately detects somatic SNVs and indels in PTA-amplified cells**

152 SCAN2 is built on SCAN-SNV, a single-cell somatic SNV genotyper that accounts for allelic  
153 imbalance (the uneven amplification of homologous alleles)<sup>14</sup>. This is achieved by measuring  
154 the VAFs of heterozygous germline SNPs, which reflect the local allelic imbalance, near  
155 candidate sSNVs. SCAN2 incorporates two key advances over SCAN-SNV. First, we developed a  
156 novel multi-sample mutation signature-based approach to increase sensitivity for sSNVs and to  
157 provide a source of information orthogonal to VAF. In short, the method relies on differences  
158 between the mutation signatures of true somatic SNVs and amplification artifacts to rescue  
159 candidate sSNVs which are rejected by the SCAN-SNV model but are poor matches to the  
160 artifact signature. The approach operates in two passes (**Fig. 2b**, Methods): in the first pass, a  
161 set of high-specificity sSNVs is produced by running SCAN-SNV in single-sample mode with  
162 stringent calling parameters. These high-specificity sSNVs are then combined across cells to  
163 generate the mutation spectrum of the true mutational process. In the second pass, candidate  
164 sSNVs rejected in the first pass are re-assessed based on their mutation contexts and  
165 potentially rescued. To do this, exposures to the learned true mutation spectrum and a  
166 universal PTA artifact signature (for derivation of this signature, see Methods and  
167 **Supplementary Fig. 3**) are computed individually for each cell; then, based on the cell-specific  
168 mutation signature exposures, each mutation context is assigned a weight representing the  
169 likelihood of originating from the artifact signature; finally, the weights are used to adjust the  
170 SCAN-SNV FDR heuristic<sup>14</sup> for rejected candidate sSNVs, allowing some candidates to be  
171 accepted (**Fig. 2c**, **Supplementary Fig. 4**). Although other multi-sample single-cell genotypers

172 exist<sup>22</sup>, our method is unique in its capability to use cross-sample information to call private  
173 sSNVs, such as those that accumulate in post-mitotic cells.

174  
175 The second key advance is the ability to call somatic indels in single-cell data. We hypothesized  
176 that, unlike artifactual sSNVs, artifactual indels are more likely to be recurrent owing to  
177 processes such as polymerase stutter<sup>23</sup> and microhomology-mediated chimera formation<sup>24</sup> that  
178 favor certain genomic regions. To identify indel artifacts, SCAN2 requires input from at least 2  
179 distinct individuals to build a list of indel sites that are frequently mutated in multiple,  
180 unrelated cells. Candidate somatic indels are initially generated by a modified SCAN-SNV  
181 protocol and then screened against the multi-subject panel to remove recurrent candidates, as  
182 they are likely artifactual (Methods, **Supplementary Fig. 5**). While this filtration proves effective  
183 at removing many indel artifacts, it is expected to limit the ability to call somatic indels at  
184 hypermutable sites that are likely to occur in many individuals such as microsatellites<sup>25</sup>.

185  
186 To assess the performance of SCAN2, synthetic diploid X chromosomes were simulated as  
187 previously described<sup>14</sup>. The multi-sample sSNV calling approach yields a mean sensitivity of  
188 45.7%, 0.0143 FPs per megabase and mean FDR of 5.9% ± 6.8% at typical somatic mutation  
189 loads for healthy cells (**Supplementary Fig. 6a-c**). Notably, the multi-sample signature approach  
190 outperformed the single-sample approach in both sensitivity and FDR at every simulated  
191 mutation burden, ranging from 0.05 sSNVs/Mb-1.5 sSNVs/Mb. Furthermore, across the same  
192 mutation burden range, multi-sample SCAN2's FDRs were lower than both Monovar<sup>21</sup> and  
193 SCcaller<sup>26</sup>, two single-cell SNV genotypers developed for MDA-amplified single cells  
194 (**Supplementary Fig. 7**). We additionally found that SCAN2 is capable of accurately predicting  
195 the total mutation burden in PTA-amplified cells by estimating and correcting for detection  
196 sensitivity using germline SNPs (Methods, **Supplementary Fig. 6d**).

197  
198 Assessment of somatic indel calling is complicated by the wide array of possible indels and the  
199 fact that indel detection sensitivity is affected by several indel characteristics, such as length  
200 and genomic context. We therefore generated a panel of indels with uniform representation  
201 across the ID83 classes, a set of 83 indel classes recently developed to enable mutation  
202 signature analysis of indels<sup>27</sup>, and used the synthetic diploid spike-in approach to score SCAN2's  
203 sensitivity separately on each of the 83 channels. SCAN2 indel sensitivity ranged from 1.4%-  
204 31%, with a clear pattern of reduced sensitivity for indels in tandem repeats greater than 4  
205 units (**Supplementary Fig. 8**). Of particular interest, we found that cross-sample filtering  
206 considerably decreased sensitivity for single base insertions in long homopolymers, which are  
207 the primary constituents of two indel aging signatures in the COSMIC catalog (ID1 and ID2). We  
208 therefore expect that correcting for ID83 class-specific sensitivity will be crucial for somatic  
209 indel signature analysis. The FP rate for somatic indels did not exceed 0.001 FPs/Mb.

210

### 211 **Revised rates of nonclonal somatic SNVs in aging human neurons**

212 SCAN2 identified 22,292 nonclonal sSNVs in the 51 MDA-amplified neurons using single sample  
213 calling and 7,174 across the 25 PTA neurons using the multi-sample approach informed by the  
214 PTA universal artifact signature. *De novo* signature extraction applied to the PTA sSNVs  
215 produced a single signature strongly resembling Signature A (cosine similarity 0.966), providing

216 confirmation of the aging-associated signature we previously recovered from MDA-amplified  
217 neurons<sup>6</sup> (**Supplementary Figure 9**). SCAN2 estimated the yearly rate of sSNV accumulation to  
218 be 14.7 sSNVs/year in PTA neurons compared to 25.7 sSNVs/year in MDA neurons from the  
219 same individuals. These rate estimates are not affected by differences in the multi-sample and  
220 single sample approaches, meaning that the difference is most likely explained by FP calls  
221 caused by greater MDA artifact burden (**Fig. 2d**) as was the case on the male X chromosomes.  
222 Nearly identical rates were produced by LiRA, a single-cell genotyper that uses an orthogonal  
223 approach both for calling sSNVs and for estimating the total sSNV burden per cell  
224 (**Supplementary Figure 10**). Importantly, although LiRA generates accurate calls based on read-  
225 level phasing, it is limited to genomic regions in close proximity to germline SNPs for phasing<sup>21</sup>;  
226 in contrast, SCAN2 can call mutations several kb from the nearest SNP and thereby generates a  
227 5-fold increase in the number of sSNV calls.

228  
229 To explore the nature of potential MDA artifacts, we focused on samples from the youngest  
230 subjects, infants, which should have the smallest true mutational burden. Amongst these  
231 samples, MDA neurons contain ~12-fold more SCAN2 sSNV calls than PTA neurons from the  
232 same individual after correcting for sensitivity, suggesting that infant MDA sSNVs can be  
233 regarded as a highly concentrated set of MDA artifacts. We first compared the infant MDA  
234 mutation spectrum with the higher quality infant PTA spectrum and found MDA sSNVs to be  
235 enriched for C>T mutations (85% vs. 59%, MDA vs. PTA) (**Fig. 2e-g**). Second, we noticed striking  
236 similarities between the infant MDA spectrum and two previously reported signatures that  
237 manifest in ways consistent with technical artifacts. Signature B (**Fig. 2h**) was previously  
238 reported in aging human neurons but did not increase with age<sup>6</sup>; Signature scF (**Fig. 2i**) was  
239 previously observed in MDA-amplified single cells but not in clonally expanded single cells from  
240 the same cell lines<sup>17</sup>. Third, we hypothesized that if these signatures are indeed artifactual, then  
241 their removal from MDA neurons would result in sSNV accumulation rates more consistent with  
242 PTA neurons. Indeed, after subtracting the Signature B-like exposure from MDA neurons, the  
243 yearly accumulation rate by SCAN2 decreased from 25.7 sSNVs/year to 16.7 sSNVs/year, more  
244 closely matching that of PTA neurons (**Supplementary Fig. 11**). Taken together, these  
245 observations provide compelling evidence that sSNVs accumulate in human neurons at a rate  
246 closer to 15 sSNVs/year and that Signature B consists largely of MDA technical artifacts.

247  
248 Finally, we emphasize that although a majority of SCAN2's calls in infant PTA neurons are C>Ts,  
249 they are materially different from those found by SCAN2 in MDA neurons and are more likely to  
250 be true mutations. This is easily seen upon computing enrichment for C>Ts by normalizing by  
251 the frequencies of NCN trinucleotide contexts in the human genome (Methods). After  
252 normalization, PTA C>Ts show a clear and strong preference for CpG contexts in a manner  
253 similar to COSMIC signature SBS1 (**Fig. 2f-i**, right panel), a mitotic clock-like signature believed  
254 to occur during cell division<sup>28</sup>. This suggests cell division during embryogenesis and subsequent  
255 development as plausible sources for infant PTA C>Ts. Among the normalized MDA spectra, a  
256 similar but smaller bias toward CpG contexts exists in the infant MDA calls and Signature B but  
257 not in Signature scF. These data suggest that neurons in the infant brain contain lower levels of  
258 single-neuron sSNVs than previously reported, but, since we remove any sSNV present in

259 matched bulk, also underestimates the number of clonal sSNVs in neurons which are likely to  
260 number in the hundreds<sup>29</sup>.

261

## 262 **Characteristics of somatic indels in single human neurons**

263 SCAN2 provides the first catalog of somatic indels from single cells and the first such catalog  
264 from a post-mitotic human cell. In total, 532 indels were identified from the 25 PTA-amplified  
265 neuronal genomes. Somatic indels increased with age by 2 to 4 somatic indels per neuron per  
266 year (Methods, **Fig. 3a**), which is surprisingly similar to rates observed in several mitotically  
267 active cell types<sup>8-10,30</sup>. However, we caution that these rates are difficult to calculate for the  
268 reasons explained above: indel sensitivity is highly dependent on indel length and genomic  
269 context and, in particular, our method has low sensitivity for highly mutable sites such as  
270 microsatellites that may recur in multiple individuals. We therefore propose a rate of ~2  
271 somatic indels per year as a lower bound. Deletions accumulated 3.3-fold faster than insertions  
272 (**Fig. 3b**) and indel sizes ranged from -28 bp to +14 bp (**Fig. 3c**). As was the case for sSNVs, MDA  
273 yields a higher accumulation rate of 3.0 somatic indels/year and we again attribute this increase  
274 to MDA artifacts; MDA somatic indels are not included in the following analyses.

275

276 Similar to sSNVs, somatic indels occur more frequently in genic regions, and the enrichment for  
277 both forms of mutation is significantly increased in highly transcribed genes (**Fig. 3d**). Of the 22  
278 exonic indels detected, 7 were scored as high impact (frame shift mutations in TIA1, MYO3B,  
279 PASK, CCDC162P, ZSCAN32, FAM161B, and CHSY1); in contrast, only 3 sSNVs were scored as  
280 high impact (stop gain in ZDHHC12, structural interaction change in PIP4K2B and a splice  
281 acceptor mutation in ANGPTL4). After adjusting for detection sensitivity, 24 high severity  
282 somatic indels and 6 high severity somatic SNVs would be expected to exist in the PTA cohort  
283 (**Fig. 3e**), suggesting that indels may have an equal or greater functional impact compared to  
284 sSNVs despite accumulating at an ~8-fold lower rate.

285

286 *De novo* mutation signature extraction yielded only a single ID83 somatic indel spectrum, likely  
287 due to the limited number of somatic indels (**Fig. 3f**), that resembles spectra from dividing  
288 cells<sup>9,10,30</sup> (**Supplementary Fig. 12**). After correcting for ID83 class-specific sensitivity, fitting to  
289 the COSMIC signature catalogue and removing signatures with <5% contribution, 7 indel  
290 signatures were detected, including two clock-like signatures ID5 and ID8 (**Fig. 3g**,  
291 **Supplementary Fig. 13**). The two remaining clock-like signatures ID1 and ID2 were not  
292 detected, consistent with the facts that neurons are post-mitotic and that the proposed  
293 aetiology for ID1 and ID2 involves DNA replication. The most prevalent signature was ID4: a  
294 signature observed in several cancer types but with no proposed mechanism. Surprisingly, ID4  
295 is more strongly correlated with age in neurons than the clock-like signatures ID5 and ID8 (**Fig.**  
296 **3h**; correlation with age = 0.86, 0.53 and 0.72, respectively). ID3 was recently detected in  
297 normal bronchial epithelium<sup>30</sup>, especially in smokers, and also shows correlation with age in  
298 neurons (correlation = 0.73). The remainder of the detected signatures (ID9, ID10 and ID11) are  
299 relatively poorly correlated with age and may represent artifacts of the signature fitting  
300 process.

301

## 302 **Discussion**

303 It is now clear that MDA genome amplification can suffer from single-stranded dropout,  
304 creating C>T artifacts that are often indistinguishable from mutations. These artifacts can be  
305 separated out by mutation signature analysis in some applications: for example, we successfully  
306 identified an sSNV signature that increases with age in human neurons despite the presence of  
307 these MDA artifacts<sup>6</sup> and confirmed this signature using PTA. Further, the similarity between  
308 SNV accumulation rates from PTA cells and MDA cells after subtracting signature B suggests  
309 that an improved correction method may be able to accurately estimate total mutation  
310 burdens from MDA. PTA introduces fewer artifacts due to its quasilinear amplification process  
311 and offers the ability to call individual mutations with high specificity. However, even using PTA,  
312 cells with low mutation burdens must be analyzed by highly specific genotypers aware of single-  
313 cell amplification artifacts.

314  
315 The methods introduced in SCAN2 come with important caveats. First, the multi-sample sSNV  
316 calling approach must be applied to batches of PTA-amplified single cells that have been  
317 exposed to similar mutational processes. Further, the efficacy of the multi-sample mutation  
318 signature approach depends on the similarity between the true signature under study and the  
319 universal PTA artifact signature: higher similarity will yield fewer benefits. The worst-case  
320 scenario occurs when the two signatures are identical; under these circumstances multi-sample  
321 calling would yield no improvement. Somatic indel detection depends on a sufficiently large  
322 sample set for screening recurrent artifacts. Notably, this filtration strategy is expected to limit  
323 SCAN2's ability to detect somatic indels at highly mutable sites such as microsatellites.

324  
325 In this study we examine indels in post-mitotic single cells for the first time. Because these cells  
326 no longer divide, the active mutational processes must not be associated with DNA replication.  
327 This may help to narrow down the possible mechanisms underlying indel signatures ID4 and  
328 ID5, whose aetiologies remain unknown. Transcriptionally associated mechanisms are the  
329 clearest candidate for further inquiry due to the enrichment of indels in expressed genes<sup>31</sup>;  
330 however, larger datasets are needed to draw conclusions with confidence.

331  
332

## 333 **Methods**

334

### 335 **Human tissue and case selection**

336 Postmortem frozen human tissues were obtained from the NIH Neurobiobank at the University  
337 of Maryland School of Medicine. Samples were obtained and processed according to IRB-  
338 approved protocol. Non-disease neurotypical individuals had no clinical history of neurologic  
339 disease and were selected to represent a range of ages from infancy to older adulthood.

340

### 341 **Isolation of single neuronal nuclei for single-cell whole genome sequencing**

342 Single neuronal nuclei were isolated using fluorescence-activated nuclear sorting (FANS) for  
343 NeuN, as described previously<sup>6,32</sup>. Briefly, nuclei were prepared from unfixed frozen human  
344 brain tissue, previously stored at -80°C, in a dounce homogenizer using a chilled tissue lysis  
345 buffer (10mM Tris-HCl, 0.32M sucrose, 3mM Mg(OAc)<sub>2</sub>, 5mM CaCl<sub>2</sub>, 0.1mM EDTA, 1mM DTT,  
346 0.1% Triton X-100, pH 8) on ice. Tissue lysates were carefully layered on top of a sucrose  
347 cushion buffer (1.8M sucrose 3mM Mg(OAc)<sub>2</sub>, 10mM Tris-HCl, 1mM DTT, pH 8) and ultra-  
348 centrifuged for 1 hour at 30,000 x g. Nuclear pellets were incubated and resuspended in ice-  
349 cold PBS supplemented with 3mM MgCl<sub>2</sub>, filtered (40 µm), then stained with Alexa Fluor 488-  
350 conjugated anti-NeuN antibody (Millipore MAB377X). Large neuronal nuclei were then  
351 subjected to FANS, one nucleus per well into 96-well plates.

352

### 353 **Single nucleus whole genome amplification by primary template-directed amplification (PTA)**

354 Isolated single neuronal nuclei were lysed and their genomes amplified using PTA, a recently  
355 developed method that pairs an isothermal DNA polymerase with a termination base<sup>15</sup>. PTA  
356 reactions were performed using the ResolveDNA EA Whole Genome Amplification Kit (formerly  
357 SkrybAmp EA WGA kit) (BioSkryb, Durham, NC), using the manufacturer's protocol. Briefly,  
358 single nuclei were sorted into wells containing 3 µL Cell Buffer pre-chilled on ice, then alkaline  
359 lysed on ice with MS Mix, mixed at 1400rpm, then neutralized with SN1 Buffer. SDX buffer was  
360 then added to the neutralized nuclei followed by a brief incubation at room temperature.  
361 Reaction-Enzyme Mix were added, then the amplification reaction was carried out for 10 hrs. at  
362 30°C, followed by enzyme inactivation at 65°C for 3 min. Amplified DNA was then cleaned up  
363 using AMPure, and yield determined by the picogreen method (Quant-iT dsDNA Assay Kit,  
364 ThermoFisher). Samples were subjected to quality control by multiplex PCR for 4 random  
365 genomic loci as previously described<sup>6</sup>, and by Bioanalyzer for fragment size distribution.  
366 Amplified genomes demonstrating positive amplification for all 4 loci were then prepared for  
367 Illumina sequencing.

368

### 369 **Library preparation for scWGS**

370 Libraries were made following a modified KAPA HyperPlus Library Preparation protocol  
371 provided in the ResolveDNA EA Whole Genome Amplification protocol. Briefly, end repair and  
372 A-tailing were performed for 500 ng of amplified DNA. Adapter ligation was then performed  
373 using the SeqCap Adapter Kit (Roche, 07141548001). Ligated DNA was cleaned up using  
374 AMPure and amplified through an on-bead PCR amplification. Amplified libraries were selected  
375 for 300-600 bp size using AMPure. Libraries were subjected to quality control using picogreen  
376 and TapeStation HS D1000 Screen Tape (Agilent PN 5067-5584) before sequencing. Single cell

377 genome libraries were sequenced on the Illumina NovaSeq platform (150bp x 2) at 30X except  
378 for subjects 1278 (HiSeq, 60X) and 1465 (NovaSeq, 60X).

379

### 380 **Single-cell amplification quality metrics**

381 Median absolute pairwise differences (MAPD) were computed by estimating copy number in  
382 bins  $CN_i$  of size 50 kb following ref. 33; subsequently,  $MAPD = \text{median}(|\log_2 CN_i -$   
383  $\log_2 CN_{i+1}|)$ . Copy number profiles in **Fig. 1** were produced using Ginkgo<sup>34</sup> with bin size 100 kb,  
384 variable binning enabled and pseudoautosomal regions masked. Allele balance distributions  
385 were computed separately for each cell by measuring single-cell VAFs at all heterozygous SNP  
386 sites used to train the SCAN2 allele balance model and then applying R's `density` function.

387

### 388 **Large somatic copy number alteration analysis**

389 Large-scale somatic CNA analysis used Ginkgo with variable bin size=1 Mb to produce a profile  
390 of normalized read counts for all bulks in PTA single cells. Large somatic CNA candidates were  
391 defined as runs of 5 or more windows  $i$  with read depth ratio  $S_{j,i}/B_i < 0.6$  or  $> 1.4$ , where  $S_{j,i}$   
392 denotes the normalized read depth in window  $i$  in single cell  $j$  and  $B_i$  is the same normalized  
393 window in the matched bulk sample. Further, somatic CNA candidates were required to have  
394 neutral copy number in the matched bulk by the same metrics. This CNA calling procedure is  
395 crude and only intended to recover very large (>5 MB) CNAs; however, these parameters  
396 successfully recovered male X chromosomes and female Y chromosomes in bulk and the large  
397 deletions observed in the PTA-amplified neuron 5823PFC-B (**Supplementary Figure 2**). Apart  
398 from 5823PFC-B, no autosomal somatic CNAs were detected by this method.

399

### 400 **Somatic mutation calling on male X chromosomes**

401 GATK HaplotypeCaller (v3.8.1) was run in joint mode across all samples (bulk, PTA and MDA) for  
402 each individual using dbSNP 147\_b37\_common\_all\_20160601 and parameters `--`  
403 `dontUseSoftClippedBases -rf BadCigar -mmq60`. Pseudoautosomal regions were  
404 not included. The resulting VCF was filtered for SNVs using GATK SelectVariants -  
405 `selectType SNP -selectType INDEL -restrictAllelesTo BIALLELIC -`  
406 `env -trimAlternates`. Somatic SNVs and indels in single cells were called separately  
407 using the following criteria: VAF > 90%, single cell depth > median(single cell depth), 0 alternate  
408 reads in bulk, bulk depth > 10 and absence from dbSNP. A set of germline SNPs and indels for  
409 estimating sensitivity was defined by sites with bulk VAF > 90%, bulk depth > median(bulk  
410 depth) and no more than 2 reference reads in bulk. For each single cell, the fraction of these  
411 sites passing the somatic filters (except for requiring 0 alternate reads in bulk and absence from  
412 dbSNP) was used as an estimate of somatic mutation sensitivity. The final estimated number of  
413 mutations was calculated by (corrected calls) = (#somatic mutations called) / (estimated  
414 sensitivity). Excess MDA calls were called per individual as the median(corrected MDA calls) -  
415 median(corrected PTA calls).

416

### 417 **sSNV false discovery rate estimation**

418 Estimated FDR curves shown in **Figure 2a** were parameterized by

419

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

420

421 Parameters used were: PTA with GATK (ref. 15), FP rate per Mb = 0.9, sensitivity = 0.8; PTA with  
 422 SCAN2 (multi-sample calling) FP rate per Mb = 0.0143, sensitivity = 0.457 (derived from  
 423 simulation experiments, see Synthetic diploid simulations). To compute the best-case scenario  
 424 for MDA, we assumed that all artifacts caused by single stranded dropout would be erroneously  
 425 identified as true SNVs and that these would be the only source of FPs. The number of single-  
 426 stranded dropout artifacts in MDA was estimated by the excess number of sSNV calls per  
 427 hemizygous X chromosome (15.9 sSNVs). To convert to FPs per diploid megabase, the excess  
 428 rate is first doubled and then divided by 152,231,524 bp, the size of chromosome X after  
 429 removing pseudoautosomal regions. This yielded a rate of 0.21 FPs per Mb, which was applied  
 430 to the whole genome. Finally, because these FPs should be called with similar sensitivity to true  
 431 mutations, there was no need to provide a sensitivity parameter for the best-case MDA  
 432 scenario since it would cancel out in the above equation.

433

#### 434 **Multi-sample somatic SNV calling procedure with SCAN2**

435 First, a set of high quality somatic SNV calls is produced for each single cell by running SCAN-  
 436 SNV in single sample mode (as described in ref. 14) with a stringent target FDR of 1%. The true  
 437 sSNV mutation spectrum is then produced by combining calls from all 25 PTA cells into a single,  
 438 raw SBS96 mutation spectrum. In general, this multi-sample combination step should only be  
 439 applied to cells exposed to the same mutational process (e.g., treatment by the same chemical  
 440 mutagen). Exposures to the true spectrum and universal PTA artifact spectrum (described  
 441 below) are computed for each single cell by least squares fitting. Weights are computed for  
 442 each cell  $i$  and rejected sSNV candidate  $j$  using a likelihood ratio

443

$$W_{i,j} = \frac{P(\text{Trinuc. context}(s\text{SNV}_{i,j}) \mid \text{True spectrum}) P(\text{True spectrum} \mid \text{cell}_i)}{P(\text{Trinuc. context}(s\text{SNV}_{i,j}) \mid \text{Artifact spectrum}) P(\text{Artifact spectrum} \mid \text{cell}_i)}$$

444

445 where  $P(\text{Trinucleotide context}(s\text{SNV}_j) \mid \text{True spectrum})$  is the component of the true  
 446 mutation spectrum corresponding to the mutation type and context of sSNV <sub>$j$</sub>  and  
 447  $P(\text{True spectrum} \mid \text{cell}_i)$  is cell  $i$ 's estimated exposure to the true mutation signature. The  
 448 same meanings apply to the artifact spectrum. Therefore,  $W_{i,j} > 1$  indicates lower likelihood of  
 449 sSNV <sub>$i,j$</sub>  being produced by the artifact process while  $W_{i,j} < 1$  indicates higher likelihood. The  
 450 weight is used to adjust a previously described heuristic<sup>14</sup> that estimates the ratio of true  
 451 mutations  $N_T$  and artifacts  $N_A$  among candidate sSNVs with similar VAF and sequencing depth as  
 452 the candidate sSNV being evaluated. This produces a multi-sample adjusted, Phred-scaled  
 453 quality score  $Q'_{i,j}$ :

454

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

455

456 where  $\alpha_{i,j}$  and  $\beta_{i,j}$  are the type I error rate and power for sSNV<sub>*ij*</sub> estimated by the pre-  
457 amplification artifact model used by SCAN-SNV (ref. 14 provides more details on this model).  
458 Finally, the rejected candidate sSNV<sub>*ij*</sub> is accepted if it was previously rejected only by the pre-  
459 amplification artifact model (i.e., passing all other criteria from ref. 14) and  $Q'_{i,j} > 20$ ,  
460 corresponding to a desired FDR of 1%. This threshold can be set by the user.

461

#### 462 **Estimation of genome-wide somatic SNV burden**

463 In addition to providing a set of sSNV calls, SCAN2 also estimates the genome-wide somatic SNV  
464 burden by estimating sSNV detection sensitivity at a subset of the high confidence,  
465 heterozygous germline SNPs (hSNPs) used to train the allele balance model. First, SCAN2  
466 calculates the distance to the nearest training hSNP for all candidate somatic SNVs and forms  
467 the distribution of these distances. The training set of germline hSNPs is then downsampled,  
468 using importance sampling, so that the distribution of distances to the nearest hSNP matches  
469 that of somatic SNV candidates. This step is necessary because the accuracy of the spatial allele  
470 balance model increases as distance to the nearest hSNP decreases. Once the downsampled set  
471 of germline hSNPs is selected, each hSNP is individually analyzed using a leave-1-out approach:  
472 the hSNP is removed from the allele balance training set, the model predicts the allele balance  
473 at the hSNP and the hSNP is then assessed using all somatic calling criteria except for dbSNP  
474 exclusion and lack of supporting reads in bulk. Only hSNPs that meet the depth requirements  
475 for somatic calling (set by the user; default: sequencing depth of the matched bulk > 10 and  
476 depth in the single cell > 5) are assessed. Among these, the fraction  $f_h$  of hSNPs passed by the  
477 somatic caller serves as an estimate of somatic sensitivity. The rate of somatic SNVs per haploid  
478 gigabase is then

$$R_{Gb} = G \frac{N_{\text{somatic}}/f_h}{2C \cdot 10^9},$$

479

480 where  $C$  is the number of diploid gigabases of the genome with sufficient sequencing depth for  
481 analysis, as specified by the user, and is collected by GATK DepthOfCoverage at base pair  
482 resolution.  $G$  is the total genome size; for **Figure 2d**,  $G=5.845$  corresponds to the number of  
483 autosomal haploid gigabases and matches ref. 6; for synthetic diploid simulations,  $G=0.3044$ ,  
484 corresponding to twice the size of the haploid, non-pseudoautosomal region of chromosome X  
485 in GRCh37. **Supplementary Figure 6d** provides an assessment of the accuracy of this estimate in  
486 simulated data with known mutation burdens.

487

#### 488 **Deriving the universal PTA artifact spectrum**

489 The universal PTA artifact spectrum was derived in 2 steps (technical details are provided in the  
490 next paragraph). First, two sets of sSNVs enriched for artifacts were extracted for each male  
491 sample (**Supplementary Fig. 3a**): (1)  $S_{X \text{ artifact}}$  from X chromosomes (male samples only) and (2)  
492  $S_{\text{Autosomal artifact}}$  from autosomal SNV candidates with VAFs consistent with expectation for pre-  
493 amplification artifacts, as determined by the local allele balance.  $S_{\text{Autosomal artifact}}$  was added  
494 because  $S_{X \text{ artifact}}$  consisted of only 190 likely artifacts, which may be insufficient to produce a  
495 high quality 96-dimensional mutation spectrum. Second, *de novo* signature extraction was  
496 performed on  $S_{X \text{ artifact}}$ ,  $S_{\text{Autosomal artifact}}$  and an additional set  $S_{\text{PASS}}$  of high quality sSNVs  
497 (**Supplementary Fig. 3b**). The high quality sSNV set provides the true mutational signature,

498 helping to prevent true mutations in  $S_{X \text{ artifact}}$  or  $S_{\text{Autosomal artifact}}$  from being assigned to the artifact  
499 signature. De novo signature extraction produced  $N=2$  signatures, as expected: one  
500 corresponding to  $S_{\text{PASS}}$  and a second corresponding to the PTA high-VAF artifact process, which  
501 became the universal PTA artifact spectrum (**Supplementary Fig. 3c**). Estimated exposures to  
502 the true and artifact spectra confirmed that the two artifact sets were highly enriched for  
503 artifacts, contrasting with the high-quality set (**Supplementary Fig. 3d**). The similarity between  
504 the PTA universal artifact signature and the MDA artifact C>T signature is notable and provides  
505 evidence that the signature is unlikely to be an overfit to this dataset.

506  
507 In more detail, X chromosome artifacts were identified from candidate SNVs produced by GATK  
508 HaplotypeCaller (as described in *Somatic mutation calling on male X chromosomes*) by requiring  
509 the SNV candidate to: (1) occur in the non-pseudoautosomal X regions, (2) have total  
510 sequencing depth  $\geq$  median(sequencing depth) of the X chromosome, (3) be supported by at  
511 least 6 alternate reads, and (4) have  $35\% \leq \text{VAF} \leq 75\%$ . Autosomal artifacts were identified by  
512 the SCAN2 allele balance consistency (ABC,  $P_{\text{true}}$ ) and pre-amplification test ( $P_{\text{artifact}}$ )  $P$ -values  
513 (see ref. 14). Briefly, large ABC  $P$ -values indicate that the candidate SNV's VAF is consistent with  
514 the locally estimated allele balance, as should be the case for a true mutation. Large pre-  
515 amplification  $P$ -values indicate that the candidate's VAF is consistent with that expected for an  
516 early-occurring artifact. Autosomal SNV candidates which fail the pre-amplification test, pass all  
517 other SCAN2 tests and for which  $P_{\text{amplification artifact}} > P_{\text{ABC}}$  were selected as autosomal artifacts.  
518  $S_{\text{PASS}}$  is the set of SNVs called by SCAN2 in single sample mode using the stringent calling  
519 parameter `--target.fdr=0.01` (i.e., PASS sSNVs). *De novo* signature extraction was  
520 performed by SigProfiler<sup>35</sup> version 2.5.1.7, as used in other *de novo* extractions. Signature  
521 channels with values  $< 10^{-4}$  were replaced by  $10^{-5}$  to prevent channels with extreme weights.

522  
523 **Somatic indel detection with SCAN2**  
524 Candidate somatic indels are initially constructed by GATK HaplotypeCaller using the same  
525 parameters as in section *Somatic mutation calling on male X chromosomes*. Somatic indels are  
526 assessed by all tests and filters applied to somatic SNVs in standard single-sample mode and an  
527 additional single-cell depth requirement of 10 reads. Notably, the allele balance model applied  
528 to candidate somatic indels is not built using germline indels; rather, the same model trained on  
529 germline hSNPs and applied to sSNVs is used for indel calling. Somatic indels passed by this  
530 process are then filtered using the cross-sample site list by requiring either: (1) reads  
531 supporting the somatic indel exist only in single cells from one individual or (2) no single cell  
532 contains more than 2 supporting reads, regardless of the number of cells and subjects in which  
533 these indel-supporting reads appear. The cross-sample list is generated by running GATK  
534 HaplotypeCaller (with the same parameters as in indel discovery) jointly on whole-genome  
535 amplified single cells from at least two individuals. Multi-sample mutation signature calling is  
536 not applied to indels, although it may be found to be beneficial with further development.

537  
538 **Synthetic diploid simulations**  
539 Synthetic X diploids (SDs), as described in ref. 14, were used to assess the performance of  
540 SCAN2. Briefly, synthetic X diploids are constructed by merging chromosome X-mapped  
541 sequencing reads from two male, independently amplified single cells. This process creates a

542 reasonably accurate amount of allelic amplification balance and amplification artifacts. In this  
543 study, 9 SDs with 30x mean depth were generated by making all pairings of the 3 PTA cells from  
544 donor 1278 and 3 PTA cells from donor 5817. The youngest donors (0.4 and 0.6 years old) were  
545 chosen to minimize the number of true somatic mutations endogenous to each X chromosome  
546 prior to adding spike-in mutations. To identify somatic SNVs endogenous to each X  
547 chromosome, GATK HaplotypeCaller was applied jointly to the SDs and the 7 PTA donor cells  
548 using the same parameters as in *Somatic mutation calling on male X chromosomes*. An  
549 additional HaplotypeCaller run using `-mmq 1` was also performed. Endogenous sSNVs were  
550 identified by applying the following hard filters: VAF=100% or VAF >= 90% with fewer than 2  
551 reference reads; depth >= 5 in the single cell, depth > 10 in the matched bulk and no mutation  
552 supporting reads in bulk in either the mapping quality 60 or mapping quality 1 runs. A single  
553 cluster of sSNVs at chrX:77471371-77471423 that appeared to be caused by clipped alignment  
554 was manually removed from the endogenous somatic mutation list. No endogenous indels  
555 were identified.

556  
557 Each SD received 20, 50, 100, 200, 500, 1000 and 2000 spike-ins, evenly split between SNVs and  
558 indels, for a total of 63 SDs. SDs with 1000 and 2000 spike-ins were used only for the rate  
559 estimation analysis presented in **Supplementary Figure 6d**. Somatic SNV spikeins were  
560 randomly generated as previously described<sup>14</sup>. Somatic indel spikein candidates were randomly  
561 generated until ~1000 candidates were obtained for each ID83 class. Indel ID83 classes were  
562 determined by first left-aligning indels by `bcftools norm` and then using  
563 `SigProfilerMatrixGenerator`<sup>36</sup> to assign ID83 status. Somatic indel spikeins were  
564 required to be at least 150 bp away from the nearest indel spikein candidate to prevent  
565 crowding in repetitive tracts and potential alignment issues caused by clustered indels. SNV and  
566 indel spikeins were not allowed to overlap. SCAN2 was run jointly on the set of 63 SDs with the  
567 same parameters used in the analysis of single neurons. Sensitivity was calculated by the  
568 fraction of successful spike-ins recovered; any SNV call not in the endogenous sSNV or spike-in  
569 sets was considered a false positive. Due to the ambiguous nature of indel representation, indel  
570 calls were considered matches to known spike-ins if either: (1) the calls matched the spike-in  
571 indel exactly or (2) the called indel was the correct length and was located exactly 1 bp away  
572 from the spike-in location.

### 573 574 **SNV calling with Monovar**

575 Monovar commit 7b47571 was downloaded and the somatic calling strategy reported  
576 previously<sup>22</sup> was mimicked as closely as possible, using scripts developed in ref. 14 (N.B., the  
577 authors provide no script for identifying somatic mutations). Single cell BAMs were input to  
578 samtools version 1.9 with options `-BQ0 -d10000 -q 40`, which was piped into the  
579 `monovar.py` script with options `-p 0.002 -a 0.2 -t 0.05 -m 2` as recommended by  
580 the authors. To determine whether SNVs were somatic or germline, samtools was run with the  
581 same options on matched bulk data. Somatic SNVs were determined by the following filters:  
582 Monovar's genotype string must not match `./.` or `0/0`; a minimum sequencing depth of 10 in  
583 the single cell with at least 3 reads supporting the mutation; at least 6 reads in bulk with no  
584 more than 1 mutation supporting read; and single cell VAF  $\geq 10\%$  for sSNVs with >100 depth or

585 VAF  $\geq$  10% for sSNVs with depth between 20 and 100. Finally, sSNVs were filtered if any other  
586 call occurred within 10 bp.

587

### 588 **SNV calling with SCcaller**

589 SCcaller version 1.1 was run as previously reported<sup>26</sup>, using scripts developed in ref. 14. BAMs  
590 were converted to pileups using samtools version 1.3.1 with the option -C50 and hSNPs were  
591 defined using dbSNP version 147 common. Single cell somatic SNVs were called by applying  
592 SCcaller's -a varcall, -a cutoff and reasoning v1.0 script in sequence with default  
593 parameters. As recommended on SCcaller's Github README, passing somatic mutations were  
594 required to have VAF > 1/8, filter status = PASS, bulk status = refgenotype and must not  
595 have been observed in dbSNP. The standard calling parameter is  $\alpha = 0.05$ , while the stringent  
596 calling parameter is  $\alpha = 0.01$ .

597

### 598 **SNV calling with LiRA**

599 LiRA version 1f4cab4 was run following instructions on Github. The joint VCF produced  
600 internally by SCAN2 (/path/to/scansnv/gatk/hc\_raw.mmq60.vcf) for each individual was  
601 supplied as the input VCF to LiRA. All samples were processed as male regardless of sex to  
602 restrict calls to the autosomes and to use a single consistent genome size for total burden  
603 estimation (LiRA accounts for the difference in genome size between males and females due to  
604 chrY). LiRA uses a genome size  $G=6.349$  for males (see *Estimation of genome-wide somatic SNV*  
605 *burden*); to restrict to autosomal extrapolation ( $G=5.845$ ) as used in all other sections and in ref.  
606 6, LiRA total SNV burden estimates were multiplied by  $5.845/6.349$ . LiRA total burden estimates  
607 retrieved from ref. 6, supplementary table S5 were not corrected in this way since they were  
608 already computed using  $G=5.845$ .

609

### 610 **Somatic SNV analysis of single human neurons**

611 MDA and PTA single neurons were analyzed by SCAN2 with identical parameters. Non-default  
612 parameters: --abmodel-chunks=4, --abmodel-samples-per-chunk=5000, --  
613 target-fdr=0.01; data resources: human reference genome GRCh37d5, SHAPEIT phasing  
614 panel 1000GP\_Phase3 and dbSNP version 147\_b37\_common\_all\_20160601. SCAN2 was run  
615 jointly on MDA and PTA cells for each subject, but subjects were analyzed in separate runs (8  
616 total SCAN2 runs corresponding to 8 subjects). Notably, even single-sample SCAN2 uses joint  
617 GATK HaplotypeCaller to create the initial set of candidate somatic SNVs, though additional  
618 information shared across cells is not used in single-sample mode. Multi-sample SCAN2 was run  
619 jointly on the SCAN2 results for all 25 PTA samples from the per-subject SCAN2 runs. sSNV  
620 accumulation rates with age were derived from a mixed-effects linear model that accounts for  
621 the fact that multiple neurons from the same individual are not independent measurements, as  
622 would be assumed by a simple linear regression. Mixed-effects model fitting was performed  
623 using the lme4 R package with the command `lmer(age ~ total_burden +`  
624 `(1|subject))`, where `total_burden` refers to the genome-wide burden estimate  
625 described in *Estimation of genome-wide somatic SNV burden*.

626

627 Mutation spectra in **Figures 2f,g** are the counts of passing sSNVs from high-confidence, single-  
628 sample SCAN2 over samples 1278BA9-A, 1278BA9-B, 1278BA9-C, 5817PFC-A, 5817PFC-B and  
629 5817PFC-C (infant PTA) and samples 1278\_ct\_p1E3, 1278\_ct\_p1E6, 1278\_ct\_p1G9,  
630 1278\_ct\_p2B9, 1278\_ct\_p2C7, 1278\_ct\_p2E4, 1278\_ct\_p2E6, 1278\_ct\_p2F5, 1278\_ct\_p2G5,  
631 5817\_ct\_p1H10, 5817\_ct\_p1H2, 5817\_ct\_p1H5 and 5817\_ct\_p2H6 (infant MDA). Multi-sample  
632 mode should not be used for mutation signature analysis since it is biased against SBS96  
633 channels that contribute to the universal artifact signature. To normalize for hg19's  
634 trinucleotide content, all 3mers (including overlaps) were extracted from the primary  
635 autosomal contigs in GRCh37 and tabulated. Each SBS96 channel was divided by the frequency  
636 of the associated 3mer in hg19.

637

### 638 **Removal of signature B from MDA samples**

639 Signature B levels in MDA samples were measured by de novo signature extraction from the  
640 combined set of 76 PTA and MDA neurons using `SigProfiler` version 2.5.1.7. 3 signatures  
641 were discovered, with one nearly identical to signature B<sup>6</sup> (cosine similarity=0.996). Removal of  
642 signature B as shown in **Supplementary Figure 11** was achieved by subtracting the reported  
643 number of sSNVs attributed to signature B from the total number of called sSNVs in each MDA  
644 sample.

645

### 646 **Somatic indel analysis of single human neurons**

647 SCAN2 was run on PTA with the same parameters used in SNV analysis (most notably, --  
648 `target.fdr=0.01`). The cross-sample filtration list was generated using all 76 MDA and PTA  
649 single cells analyzed in this study. Indels were classified into ID83 channels using  
650 `SigProfilerMatrixGenerator`. For MDA-amplified neurons only, somatic indels were  
651 additionally filtered to remove all single base insertions in homopolymers of length 3 or greater  
652 (i.e., ID83 classes 1:Ins:C: 3-5 and 1:Ins:T:3-5). Somatic indel sensitivity was computed in two  
653 ways following the process in *Estimation of genome-wide somatic SNV burden*. First, germline  
654 heterozygous indels discovered in bulk were downsampled to match the ID83 spectrum of  
655 called somatic indels to provide a set of indels with roughly similar characteristics to somatic  
656 indels. Total sensitivity was computed on the downsampled germline set, giving a sensitivity-  
657 adjusted  $N_{\text{somatic}} = (\# \text{ called indels}) / (\text{germline sensitivity})$ . Since the cross-subject panel was not  
658 applied to the germline heterozygous indels (because they are common polymorphisms and are  
659 often shared), this overestimates sensitivity and underestimates of the number of indels.  
660 Second, the ID83 channel-specific sensitivities derived from SD simulations were applied to  
661 each single cell individually by dividing the number of somatic indels calls per channel by the  
662 channel-specific sensitivity. Summing over all ID83 classes gives  $N_{\text{somatic}}$  per cell. The final rate  
663  $R_{\text{Gb}}$  was estimated as explained above. De novo extraction was performed by `SigProfiler`  
664 on PTA neurons only, which produced only a single signature. Fits to COSMIC indel signatures  
665 were performed using the COSMIC version 3 set of indel signatures ID1-18. For the discovery of  
666 active signatures in **Figure 3g**, all 532 indels were combined into a single set and exposures to  
667 each of the 17 signatures were estimated by least squares fitting using `lsqnonneg` from the  
668 `pracma` R package. Otherwise (e.g., for analysis of correlation with age), somatic indels were  
669 kept separate and fit using the same method.

670

## 671 **Functional impact of point mutations**

672 The severity of somatic SNV and indel mutations reported in Figure 3 were derived from SnpEff  
673 version 4.3t using the hg19 database. High and moderate mutations were those annotated as  
674 HIGH or MODERATE, respectively, in the first reported annotation field. The genes impacted  
675 and protein-altering effects were also taken from the first annotation field. The extrapolation  
676 from called mutations to the expected number over the PTA cohort used cohort-wide  
677 sensitivity estimates of 47.7% for sSNVs and 29.3% for somatic indels corresponding to (#  
678 SCAN2 PTA sSNV calls = 7174) / (sum of estimated PTA sSNV burdens = 15,030) and (# SCAN2  
679 PTA indel calls = 532) / (germline sensitivity-based estimate for total indel burden = 1812),  
680 respectively. The number of expected high-impact mutations per cohort is the number of  
681 observed HIGH impact mutations (n=7, indels, n=3, sSNVs) divided by sensitivity.

682

## 683 **Enrichment of somatic mutations in transcribed genes**

684 Gene expression quantification data were obtained from the GTEx consortium (version 8); gene  
685 start and stop genomic positions were obtained by matching GENCODE v26 (hg38 to b37  
686 liftover) “gene” records (column 3) to the GTEx expression matrix using Ensembl gene IDs.  
687 Autosomal genes with mean TPM > 1 in either 209 frontal cortex (BA9) GTEx samples or across  
688 the full GTEx dataset were retained for analysis. Genes retained by this filtration were then  
689 ranked by average TPM across the 209 BA9 samples and separated into expression quintiles.  
690 Each somatic mutation was assigned to 1 of 6 bins (5 expression quintiles and intergenic) based  
691 on overlap with this gene set. Mutations overlapping multiple genes were resolved by assigning  
692 the mutation to the first gene in the overlap list. Enrichment analysis was performed by  
693 permutation: for each single cell, mutation positions were randomly shuffled across the  
694 genome 250 times to create a null distribution of mutation density. To approximate calling  
695 sensitivity, position shuffling was restricted to the subset of each single cell genome that met  
696 the minimum depth requirements for SCAN2 analysis. To perform enrichment calculations,  
697 observed mutation counts for each bin  $i$  were combined across all samples for either the  
698 observed data  $D_i$  (7,174 sSNVs by multi-sample SCAN2 or 532 indels, separately) or one of the  
699 shufflings  $R_i^{(j)}$ ,  $j = 1$  to 250. Enrichment levels were calculated for observations and  
700 permutations by dividing each bin count by the mean count over the 250 shufflings  $M_i =$   
701  $\frac{1}{250} \sum_{j=1}^{250} R_i^{(j)}$ . Two-tailed  $p$ -values were determined for each bin  $i$  by counting the fraction of  
702 permutations with absolute log ratios exceeding the observed absolute log ratio:

703

$$P_i = \frac{1}{250} \sum_{j=1}^{250} I \left\{ \left| \log_2 \frac{R_i^{(j)}}{M_i} \right| \geq \left| \log_2 \frac{D_i}{M_i} \right| \right\},$$

704

705 where  $I(\cdot)$  is the indicator function.

706

## 707 **Data availability**

708 All MDA-amplified single neurons and matched bulks listed in **Supplementary Table 2** were  
709 downloaded from dbGaP, identifier phs001485.v1.p1. Only neurons from the pre-frontal  
710 cortices from individuals for which additional PTA data were generated were used. *Raw*  
711 *sequencing read data for PTA-amplified single cells will be uploaded to dbGaP.*

712

713 **Code availability**

714 SCAN2 is available for download at <https://github.com/parklab/SCAN2>.

715

716 **Competing interests.** The authors declare the following competing interests: C. G. is Director  
717 and cofounder and J. W. is CEO and cofounder of Bioskryb, Inc., the manufacturer of PTA kits  
718 used in this study.

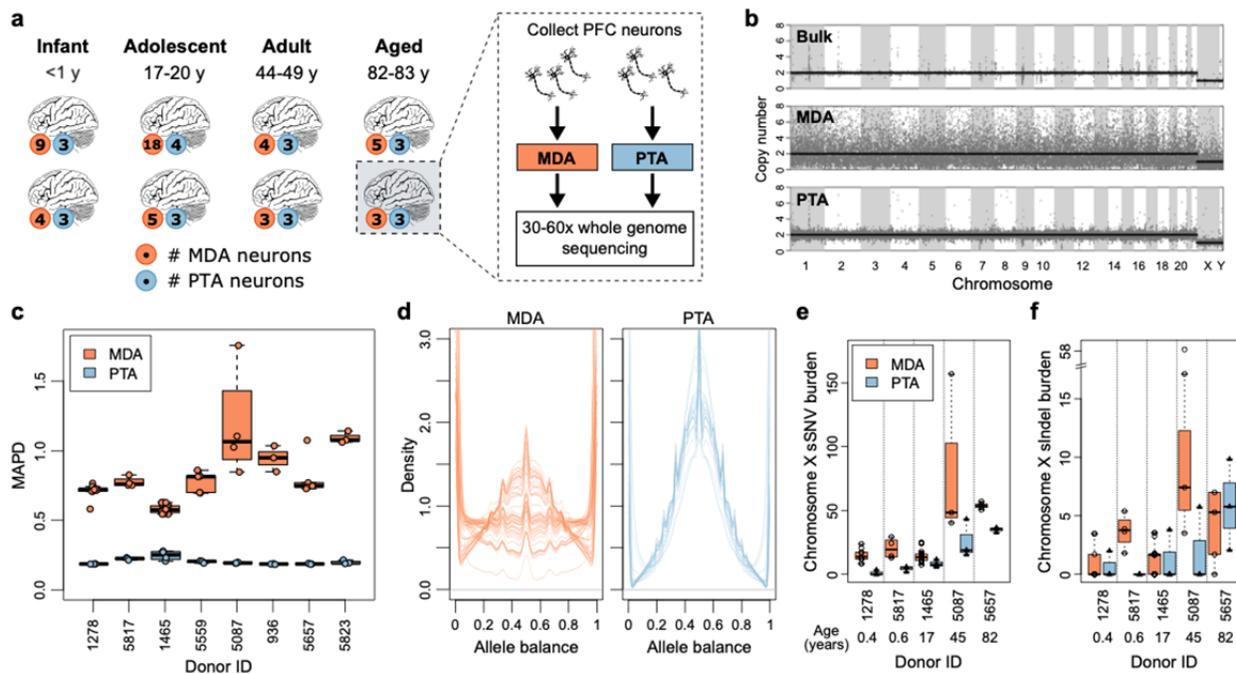
## 719 References

- 720 1. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic Mutation, Genomic Variation,  
721 and Neurological Disease. *Science* **341**, 43-51 (2013).
- 722 2. Lodato, M. *et al.* Somatic mutation in single human neurons tracks developmental and  
723 transcriptional history. *Science* **350**, 94-98 (2015).
- 724 3. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic  
725 mutations in normal human skin. *Science* **348**, 880-886 (2015).
- 726 4. Jaiswal, S. *et al.* Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease.  
727 *N. Engl. J. Med.* **377**, 111-121 (2017).
- 728 5. Blokzijl, F., de Ligt, J., Jager, M. *et al.* Tissue-specific mutation accumulation in human  
729 adult stem cells during life. *Nature* **538**, 260–264 (2016).
- 730 6. Lodato, M. *et al.* Aging and neurodegeneration are associated with increased mutations  
731 in single human neurons. *Science* **359**, 555-559 (2018).
- 732 7. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age.  
733 *Science* **362**, 911-917 (2018).
- 734 8. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells.  
735 *Nature* **574**, 532-537 (2019).
- 736 9. Franco, I. *et al.* Somatic mutagenesis in satellite cells associates with human skeletal  
737 muscle aging. *Nat Commun* **9**, 800 (2018).
- 738 10. Franco, I., Helgadottir, H. T. *et al.* Whole genome DNA sequencing provides an atlas of  
739 somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type.  
740 *Genome Biology* **20**, 285 (2019).
- 741 11. Woodworth, M. B., Girsakis, K. M., & Walsh, C. A. Building a lineage from single cells:  
742 genetic techniques for cell lineage tracking. *Nat Rev Genet* **18**, 230-244 (2017).
- 743 12. Evrony, G., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain  
744 using single-neuron genomics. *eLife* **5**, e12966 (2016).
- 745 13. Zhang, C. Z., Adalsteinsson, V.A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K.L.,  
746 Meyerson, M. & Love, J.C. Calibrating genomic and allelic coverage bias in single-cell  
747 sequencing. *Nat Commun* **6**, 6822 (2015).
- 748 14. Luquette, L. J. *et al.* Identification of somatic mutations in single cell DNA-seq using a  
749 spatial model of allelic imbalance. *Nat Commun* **10**, 3908 (2019).
- 750 15. Gonzalez-Pena, V., Natarajan S. *et al.* Accurate Genomic Variant Detection in Single Cells  
751 with Primary Template-Directed Amplification. *bioRxiv*, doi:  
752 <https://doi.org/10.1101/2020.11.20.391961>.
- 753 16. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,  
754 415-421 (2013).
- 755 17. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines  
756 Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294 (2019).
- 757 18. Ning, L., Li, Z., Wang, G., *et al.* Quantitative assessment of single-cell whole genome  
758 amplification methods for detecting copy number variation using hippocampal neurons.  
759 *Sci Rep.* **5**, 11415 (2015).
- 760 19. McConnell, M. J., Lindberg, M. R., Brennand, K. J., *et al.* Mosaic copy number variation in  
761 human neurons. *Science* **342**, 632-637 (2013).

- 762 20. Chronister, W. D., Burbulis, I. E., Wierman, M. B., *et al.* Neurons with Complex  
763 Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep* **26**, 825-835 (2019).
- 764 21. Bohrsen, C. *et al.* Linked-read analysis identifies mutations in single-cell DNA sequencing  
765 data. *Nat Genet* **51**, 749-754 (2019).
- 766 22. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant  
767 detection in single cells. *Nat Meth* **13**, 505-507 (2016).
- 768 23. Gymrek, M. PCR-free library preparation greatly reduces stutter noise at short tandem  
769 repeats. *bioRxiv* doi: 10.1101/043448 (2016).
- 770 24. Lasken, R. S., Stockwell, T. B. Mechanism of chimera formation during the Multiple  
771 Displacement Amplification reaction. *BMC Biotechnology* **7**, doi:10.1186/1472-6750-7-  
772 19 (2007).
- 773 25. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**,  
774 435-445 (2004).
- 775 26. Dong, X., Zhang, L., Milholland, B., *et al.* Accurate identification of single-nucleotide  
776 variants in whole-genome-amplified single cells. *Nat Methods* **14**, 491-493 (2017).
- 777 27. Alexandrov, L. B., Kim, J., Haradhvala, N. J. *et al.* The repertoire of mutational signatures  
778 in human cancer. *Nature* **578**, 94–101 (2020).
- 779 28. Alexandrov, L.B., Jones, P., Wedge, D. *et al.* Clock-like mutational processes in human  
780 somatic cells. *Nat Genet* **47**, 1402–1407 (2015).
- 781 29. Bae, T., Tomasini, L., Mariani, J., *et al.* Different mutational rates and mechanisms in  
782 human cells at pregastrulation and neurogenesis. *Science* **359**, 550-555 (2018).
- 783 30. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial  
784 epithelium. *Nature* **578**, 266-272 (2020).
- 785 31. Alt, F.W., Schwer, B. DNA double-strand breaks as drivers of neural genomic change,  
786 function, and disease. *DNA Repair* **71**, 158-163 (2018).
- 787 32. Evrony, G.D., Cai, X., Lee, E., *et al.* Single-neuron sequencing analysis of L1  
788 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).
- 789 33. Baslan, T., Kendall, J., Rodgers, L., *et al.* Genome-wide copy number analysis of single  
790 cells. *Nat Protoc* **7**, 1024-1041 (2012).
- 791 34. Garvin, T., Aboukhalil, R., Kendall, J., *et al.* Interactive analysis and assessment of single-  
792 cell copy-number variations. *Nat Methods* **12**, 1058-1060 (2015).
- 793 35. Ludmil Alexandrov (2020). SigProfiler  
794 (<https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler>), MATLAB  
795 Central File Exchange. Retrieved January 1, 2020.
- 796 36. Bergstrom, E. N., Huang, M. N., Mahto, U. *et al.* SigProfilerMatrixGenerator: a tool for  
797 visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685  
798 (2019).

799  
800

801



802

803

804

805

806

807

808

809

810

811

812

813

814

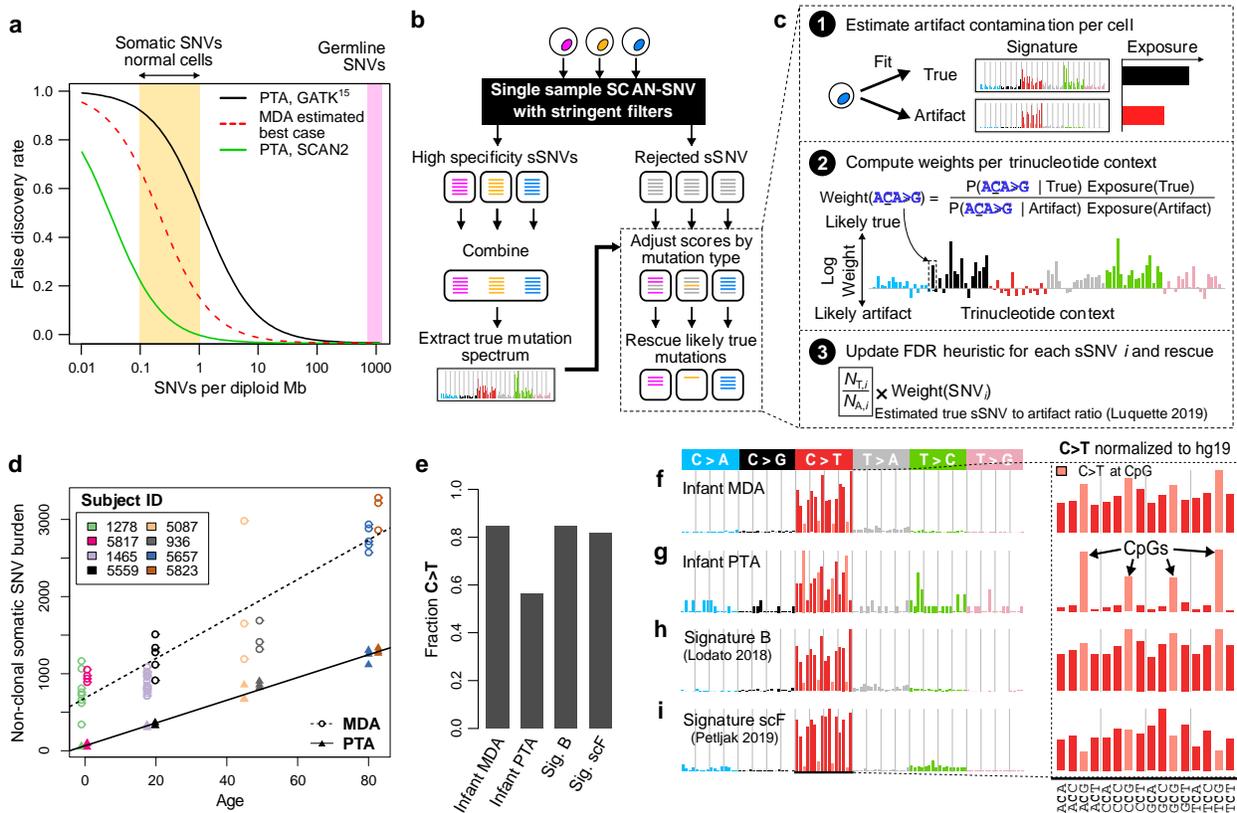
815

816

817

**Figure 1. PTA improves over MDA at large and small scales.** **a.** Study design. Single neurons were collected from the prefrontal cortices of brains of 8 individuals ranging in age from infantile to elderly. Single neurons were amplified by either PTA or MDA and then sequenced to high coverage. **b.** Representative copy number profiles for bulk (top), MDA-amplified (middle) and PTA-amplified (bottom) genomes. **c.** MAPD (median absolute pairwise deviation) 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. **d.** Allele balance for germline heterozygous SNPs in each sample. Each line corresponds to one single cell. Values near 0.5 indicate balanced amplification of homologous alleles; values near 0 or 1 indicate complete dropout of one allele. **e.** Sensitivity-adjusted somatic SNV (sSNV) burdens per X chromosome for 5 male individuals. **f.** Same as (e) for somatic indels (sIndels). Boxplot whiskers, furthest point at most 1.5x interquartile range.

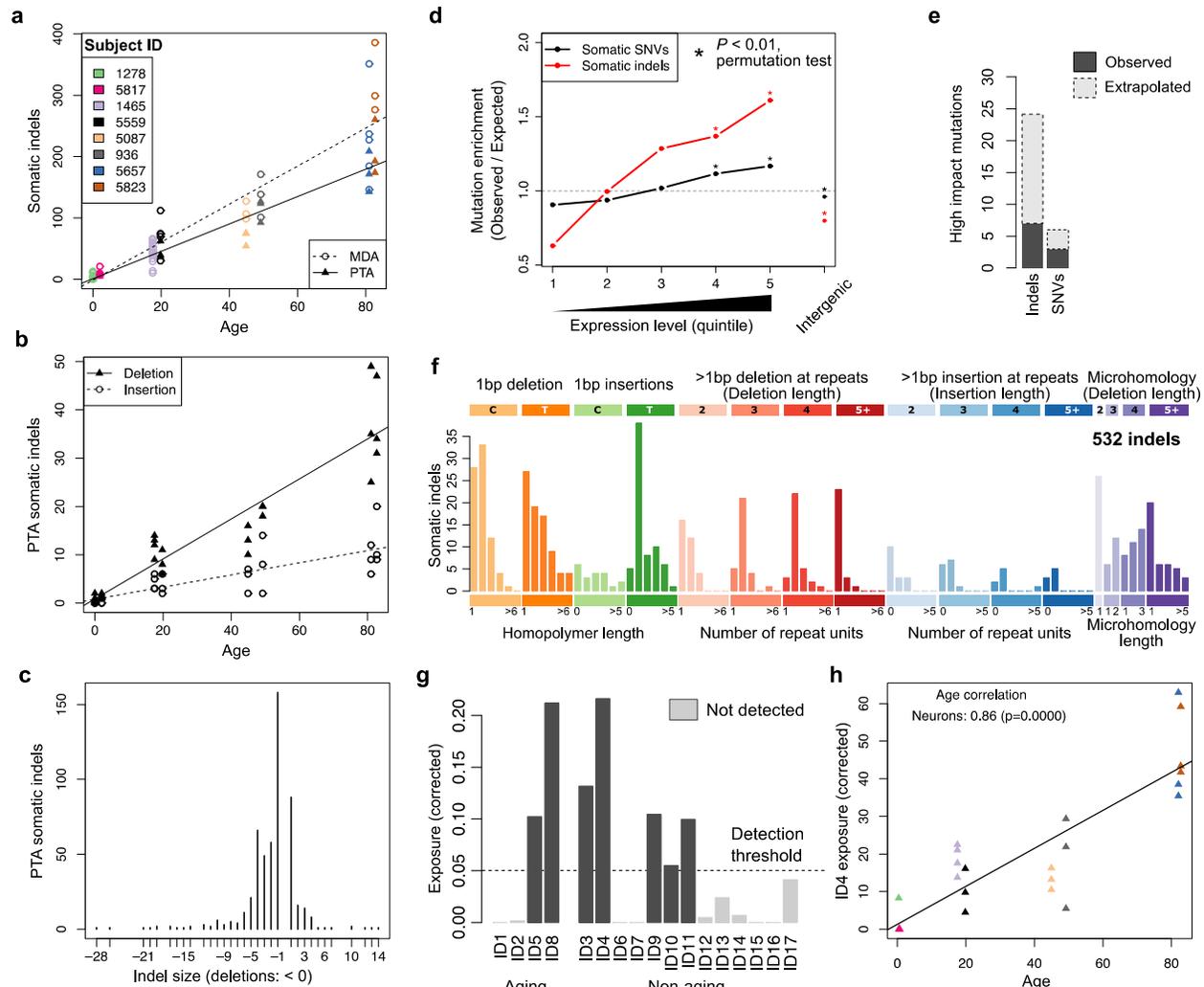
818



819  
820

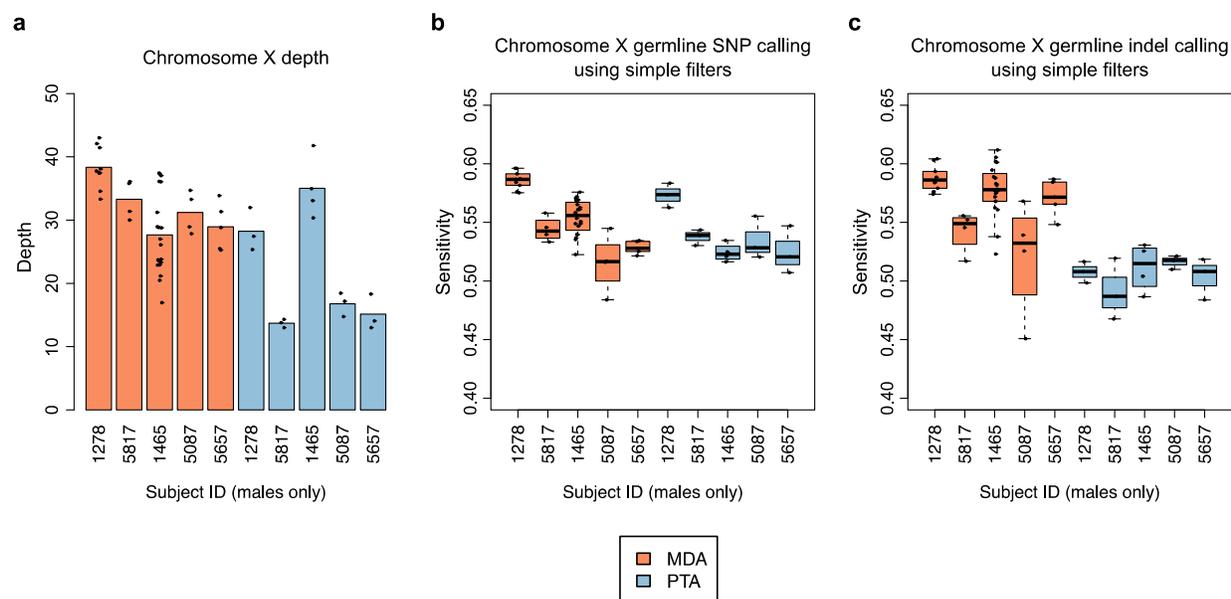
**Figure 2: SCAN2 somatic SNV calling method and application to single human neurons.**

821 **a.** Estimated false discovery rates for MDA and PTA somatic SNV detection. The MDA best  
 822 case scenario assumes that all single strand-dropout artifacts are erroneously called as true  
 823 mutations. **b.** SCAN2 approach to multi-sample sSNV calling. SCAN2's multi-sample approach  
 824 is not phylogenetic and does not depend on sSNVs being shared by multiple single cells. It can  
 825 therefore detect private mutations such as those in post-mitotic neurons. **c.** Candidate sSNVs  
 826 are rescored separately for each single cell given the true mutation signature learned in panel  
 827 (b). The likelihood of being generated by the true signature is computed for each mutation type  
 828 and trinucleotide context (x-axis). This likelihood acts as a prior for a previously described  
 829 heuristic that estimates the number of true mutations ( $N_{T,i}$ ) and artifacts ( $N_{A,i}$ ) with characteristics  
 830 similar to the sSNV candidate  $i$ . **d.** Sensitivity-adjusted accumulation rate of somatic SNVs in  
 831 PTA- (triangles) and MDA- (circles) amplified single human neurons. **e.** Fraction of C>Ts  
 832 among sSNVs called by single sample SCAN2 in infant neurons and two previously published  
 833 signatures. **f.** Mutational spectra of somatic SNVs called by SCAN2 in single-sample mode  
 834 across 6 MDA neurons from 2 infants. Signature B is not subtracted from MDA calls. Right:  
 835 rate of C>T mutations after normalizing by trinucleotide frequency in the human genome. **g.** Same  
 836 as (f) for 6 PTA neurons from 2 infant donors. **h.** C>T rich neuron signature B reported in  
 837 Lodato et al, 2018. **i.** MDA artifact signature scF reported by Petljak et al, 2019.



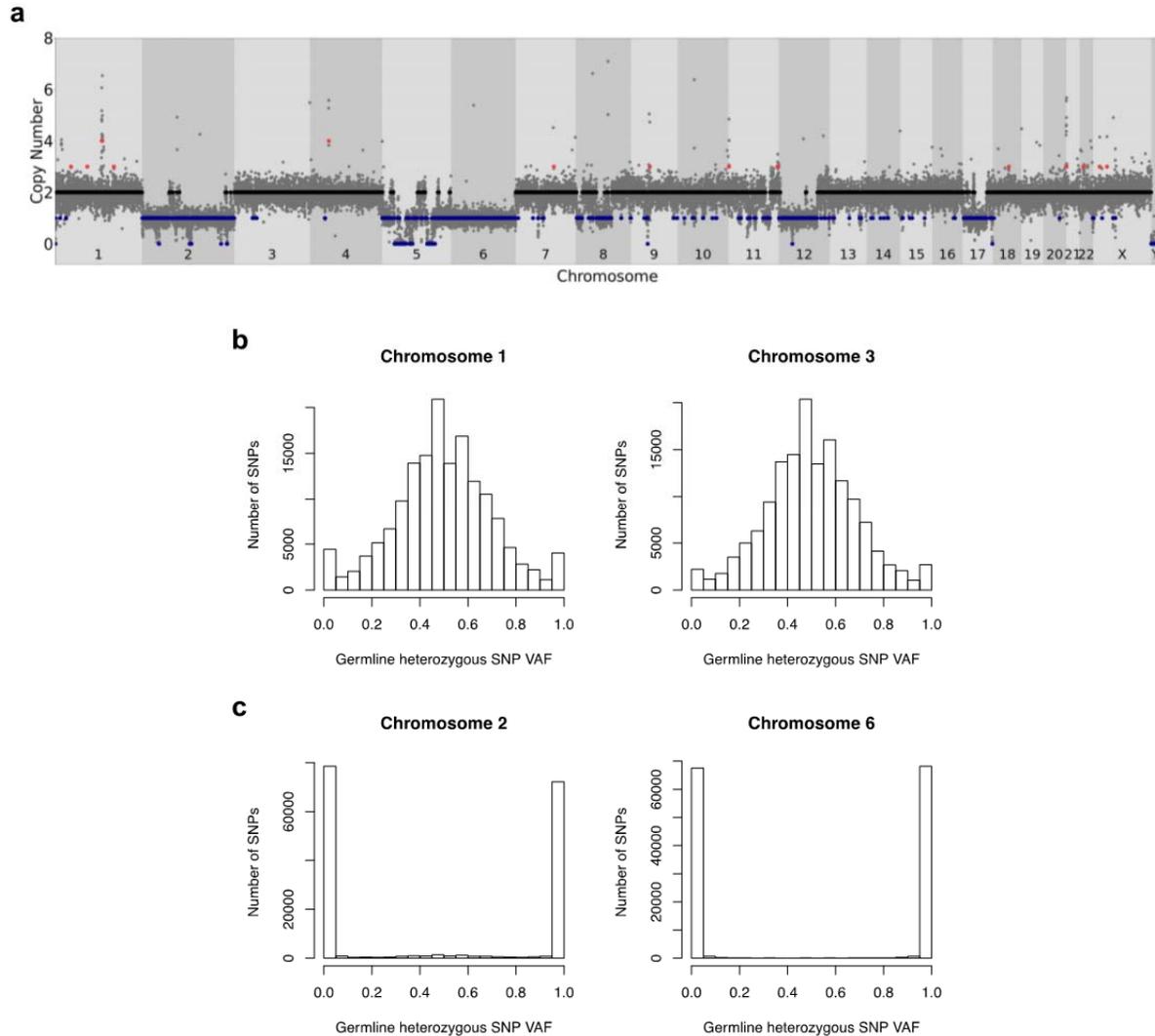
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856

**Figure 3: Characteristics of somatic indels in aging human neurons.** **a.** Age-related increase of somatic indel burden across 8 individuals. Adjustment for sensitivity shown here represents the lower bound corresponding to ~2 somatic indels per neuron per year. **b.** Age-related increase of somatic insertions and deletions called from PTA neurons, separately. **c.** Distribution of somatic indel lengths from PTA neurons. **d.** Enrichment of PTA somatic SNVs in intergenic regions and transcribed regions stratified by expression quartile. Expression levels were derived from GTEx and range from quintile 1 (lowest) to quintile 5 (highest). Expected number of mutations determined by permutation testing (\*:  $p < 0.01$ ). **e.** Number of high impact mutations according to SnpEff (dark grey); expected number of high impact mutations after adjusting for sensitivity (light grey). **f.** Mutation spectrum identified by de novo signature extraction from 532 somatic indels. **g.** Exposures to COSMIC ID signatures calculated by least squares fitting. Exposures were corrected by normalizing indel counts by ID83 channel-specific sensitivity (**Supplementary Fig. 8c**) before fitting. **h.** Age association of ID4, a signature of unknown aetiology, with neuron age.



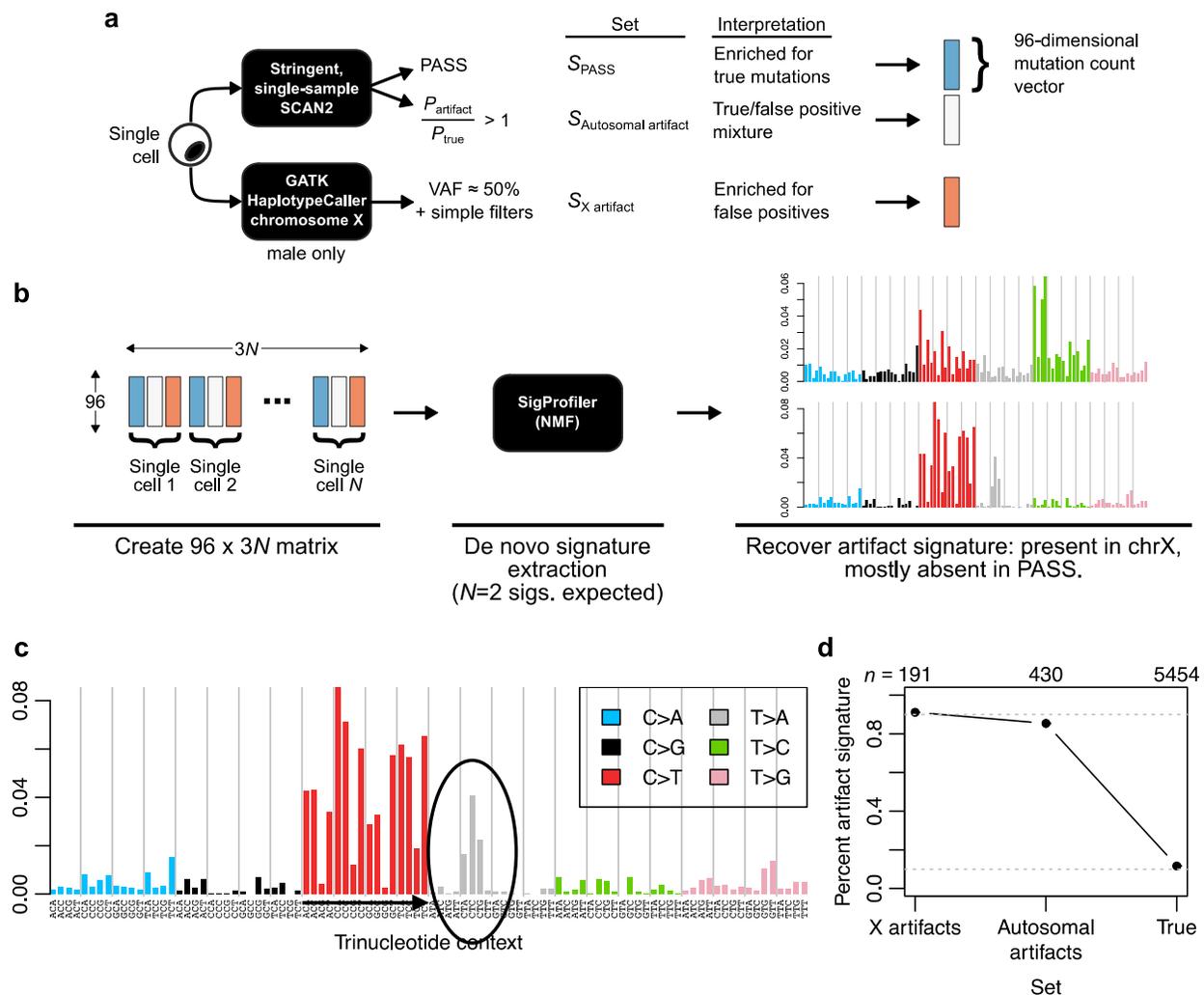
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869

**Supplementary Figure 1: Simple somatic mutation calling on male chromosome X. a.** Mean sequencing depth per cell (points) and averaged over all cells per donor (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 sequenced to about half of the genome-wide mean depth due to hemizyosity. **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 indels. Boxplot whiskers, furthest point at most 1.5x interquartile range.



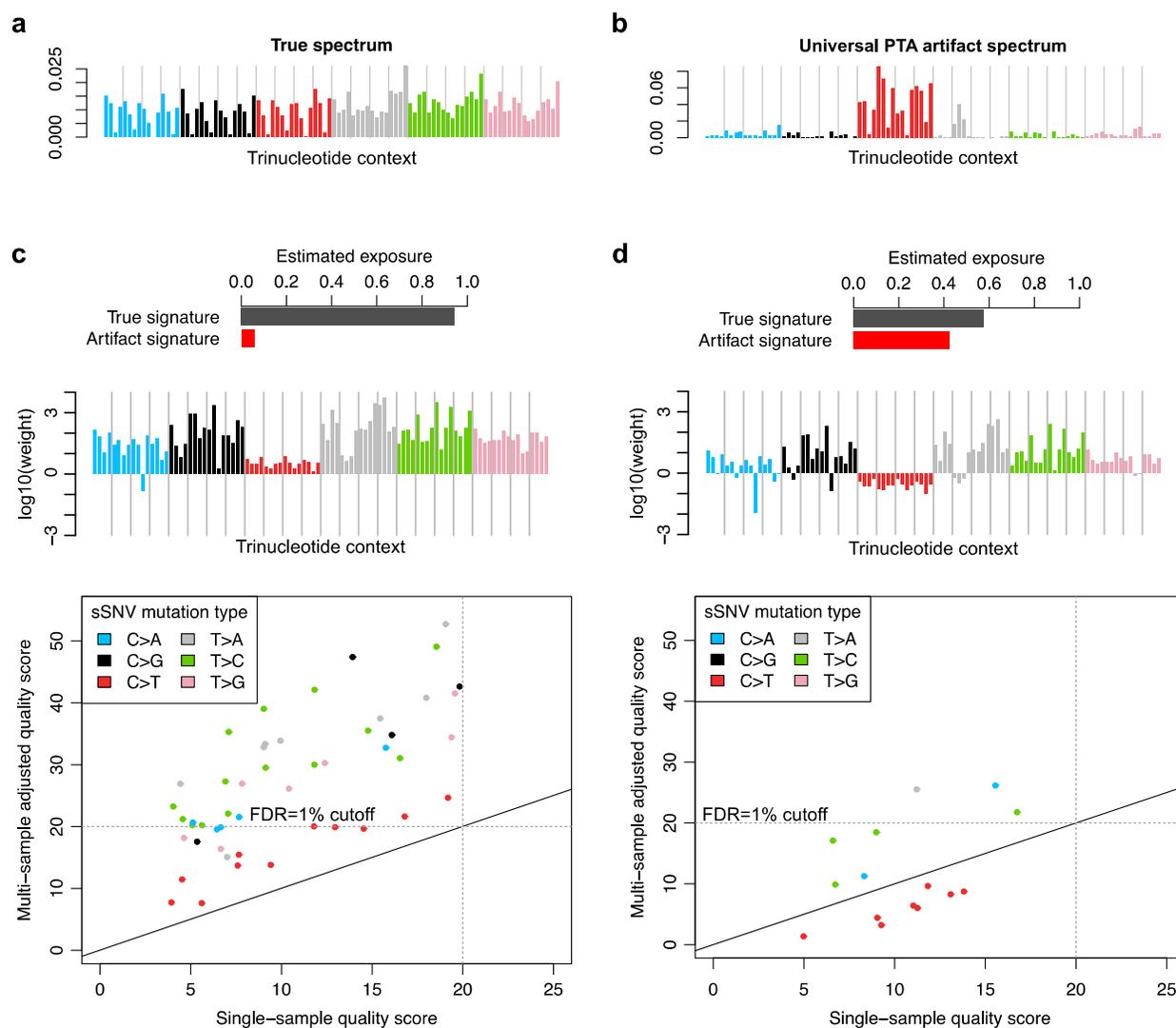
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883

**Supplementary Figure 2: Possible failed PTA amplification.** **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 single neuron truly contained a single copy loss or if the apparent loss resulted from complete amplification failure of one haplotype cannot be determined.



884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898

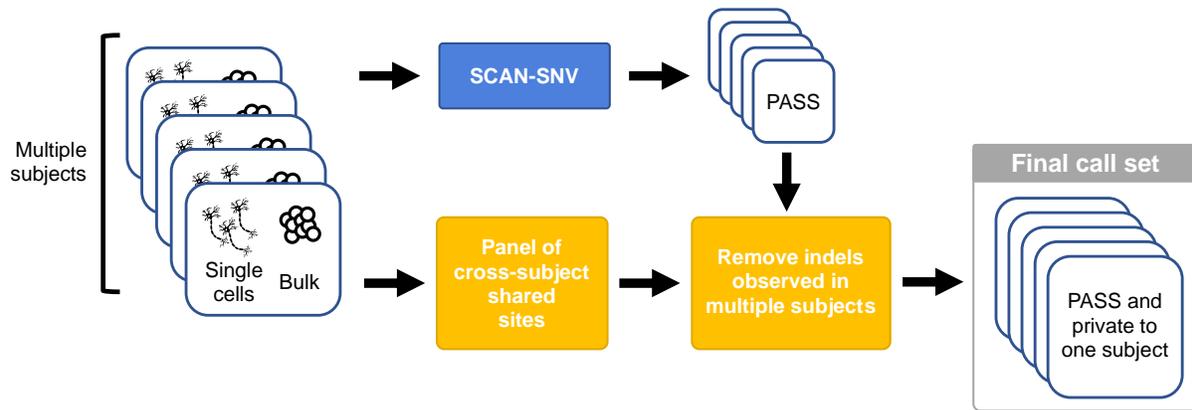
**Supplementary Figure 3: The universal PTA artifact signature.** **a.** 3 sets of SNVs and likely artifacts were constructed for each male single cell. PASS autosomal SNVs using stringent calling filters are highly depleted for artifacts while rejected candidate SNVs with  $P_{\text{artifact}}/P_{\text{true}} > 1$  (see ref. 14 for information on the models corresponding to these  $P$ -values) or chromosome X sites in the non-pseudoautosomal regions with  $\sim 50\%$  VAF in male samples are highly enriched for early, high-VAF PTA artifacts. **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<sup>6</sup> and the universal PTA artifact signature. **c.** The 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%.



899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916

**Supplementary Figure 4: Examples of the multi-sample sSNV approach: weight calculation and quality score adjustment.** **a.** True mutation spectrum derived from high confidence calls in simulated data (synthetic diploids, see **Supplementary Figure 6** for a detailed performance comparison). **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,  $y=x$ .

917



918

919

920 **Supplementary Figure 5: Somatic indel calling strategy.** PTA-amplified cells and matching

921 bulk samples from multiple subjects are required for indel calling. Single cells and bulks are

922 each analyzed by a modified SCAN-SNV pipeline. GATK HaplotypeCaller is independently run

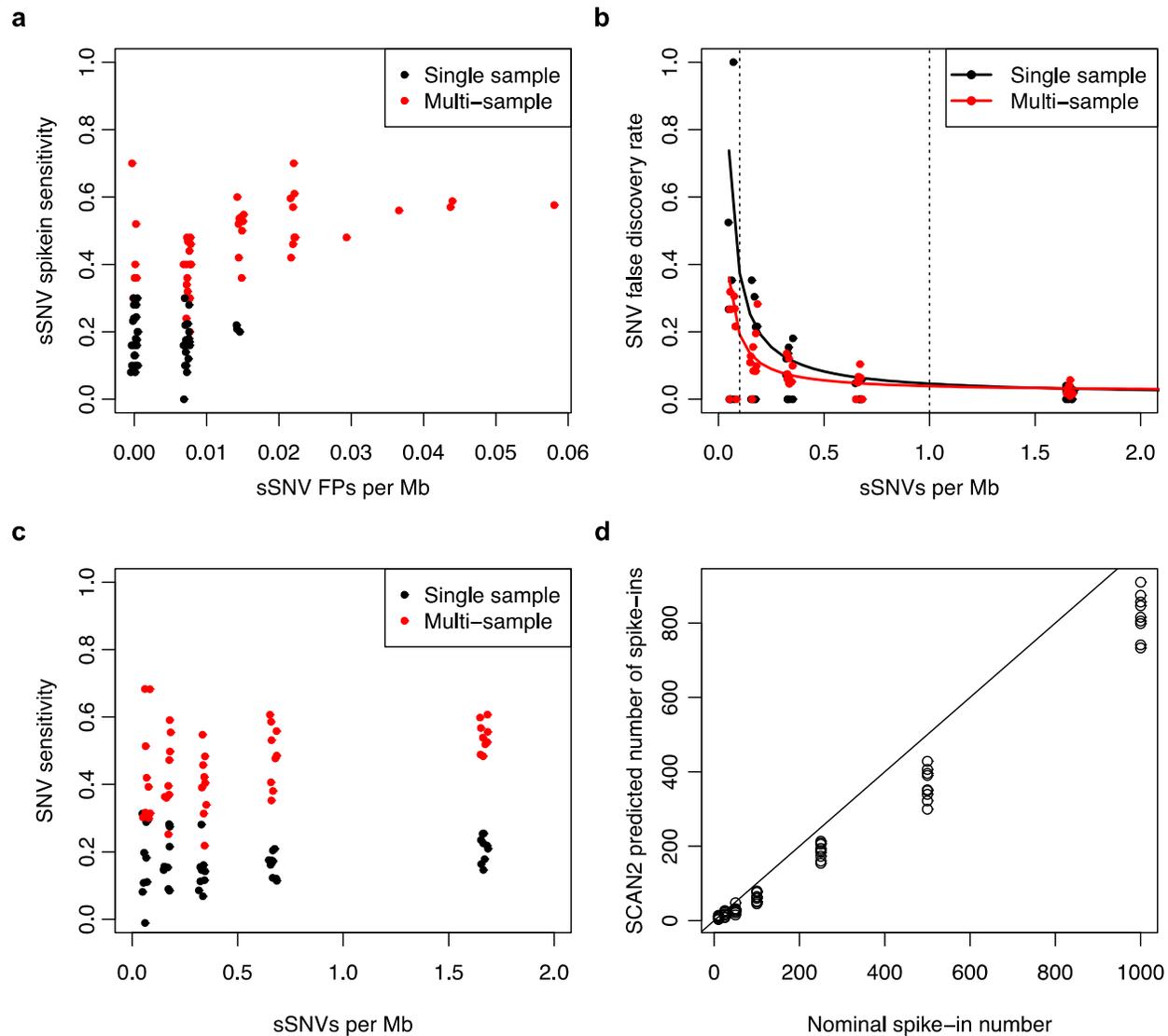
923 in joint mode on all single cells and bulks to produce a panel containing reference and alternate

924 read counts across the full cohort. Somatic indels passed by the modified SCAN-SNV pipeline

925 are then removed if reads supporting the indel are observed in single cells from other subjects.

926

927



928

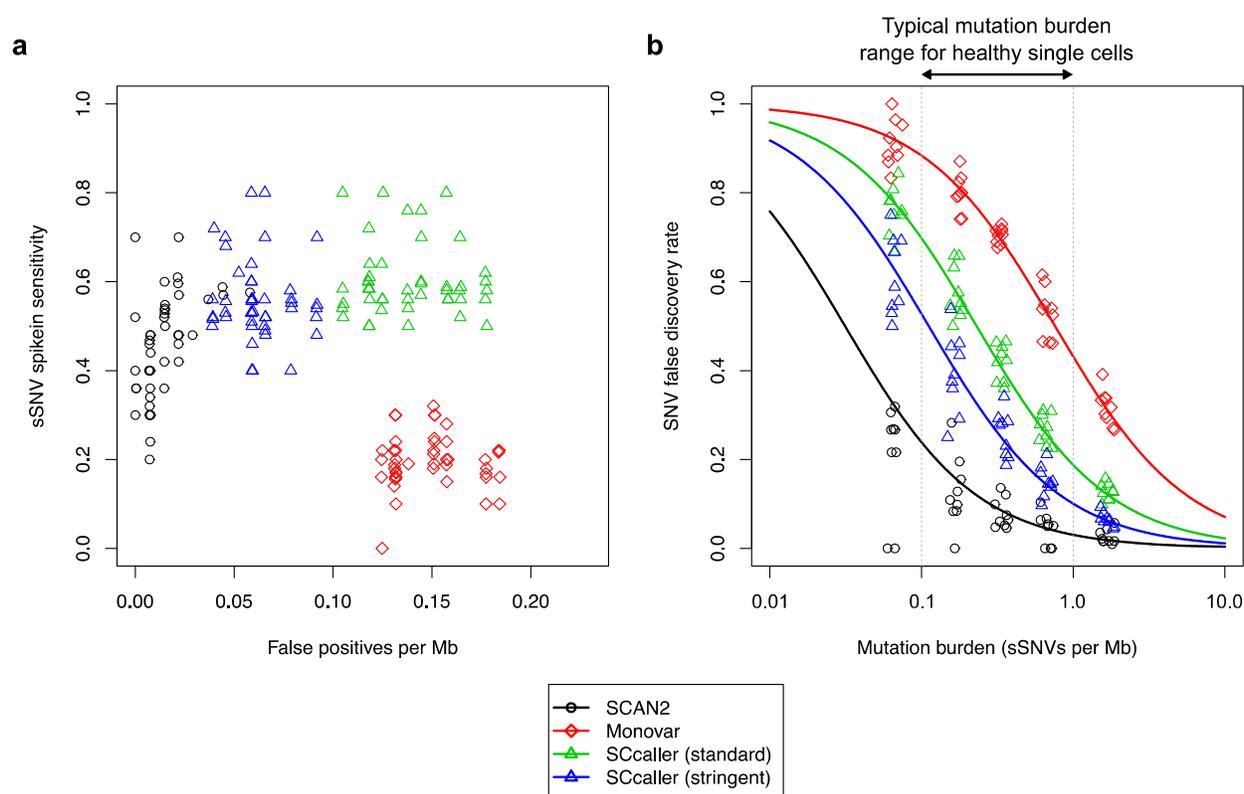
929

930 **Supplementary Figure 6: Simulated data to assess sSNV calling performance for single**  
 931 **and multi-sample SCAN2** **a.** SNV sensitivity and false positive rate for synthetic diploid  
 932 simulations with 1-250 spike-ins per simulation. Target FDR=1%, rescue FDR=1%. **b.** SNV  
 933 sensitivity plotted against mutation burden for simulated SNVs. **c.** False discovery rate plotted  
 934 against mutation burden for simulated SNVs. Solid lines: linear regression fits to  $FDR \sim$   
 935  $1/\text{mutations per Mb}$ . Dotted vertical lines: typical range of somatic mutation burdens in healthy  
 936 single cells. **d.** SCAN2 total sSNV burden estimates for 63 simulations. 9 synthetic diploid  
 937 simulations were performed for each of the spike-in rates of 10, 25, 50, 100, 250, 500 and 1000  
 938 per simulation. Solid line:  $y=x$ . x-axes for panels **a-c** are jittered for visibility.

939

940

941



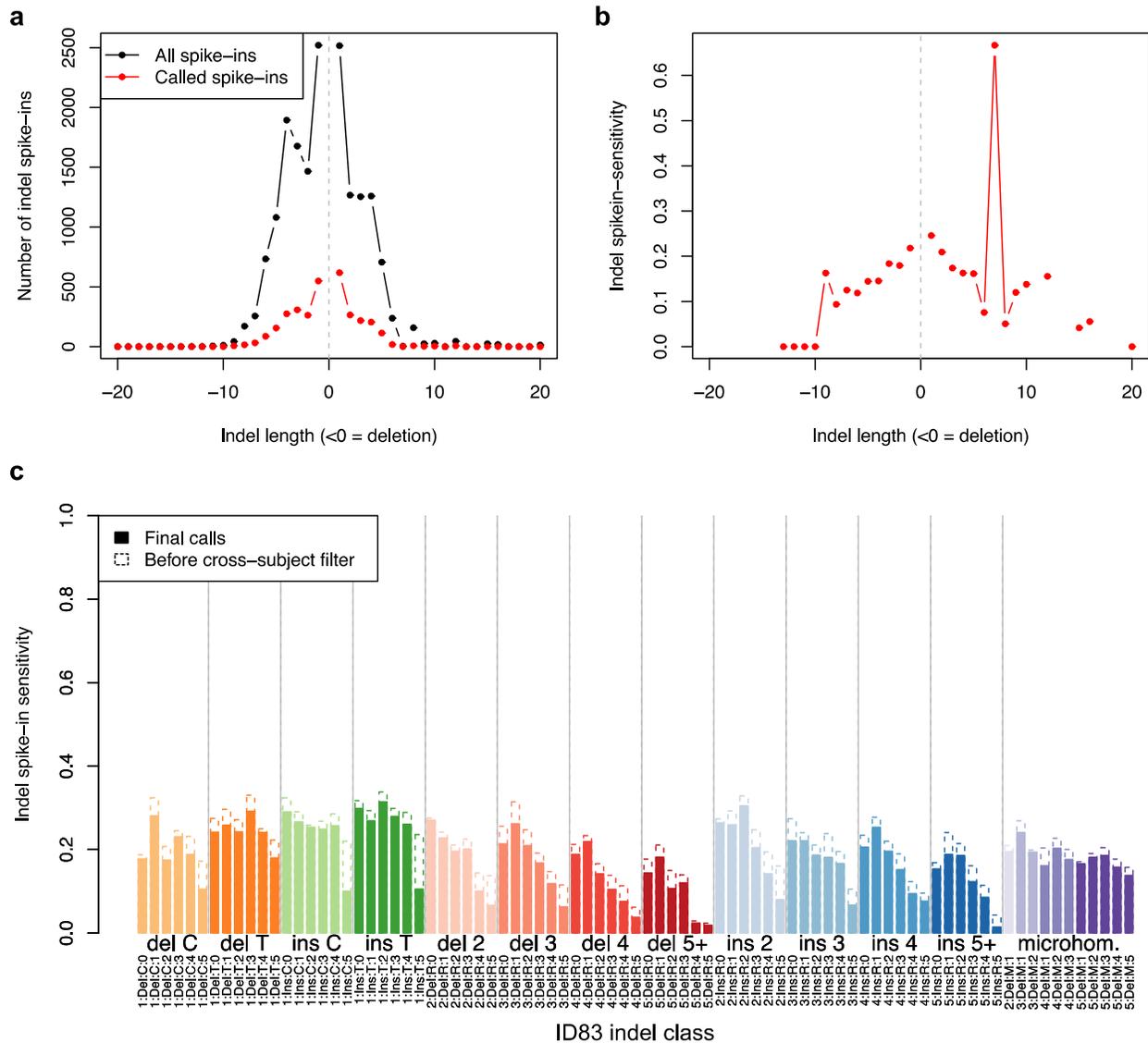
942

943

944 **Supplementary Figure 7: Comparison of SCAN2 to other single-cell SNV genotypers. a.**

945 Each point represents a single simulated synthetic diploid X chromosome. Sensitivity is the  
 946 fraction of spike-ins recovered. False positives are SNV calls that were not known spike-ins or  
 947 endogenous somatic mutations. **b.** False discovery rate vs. the number of spike-ins per  
 948 megabase. Lines are parameterized by mean sensitivity  $S$  and false positive rate per megabase  
 949  $F$ :  $FDR = F / (F + xS)$ . Single cells from non-neoplastic human tissues typically exhibit SNV  
 950 burdens between 0.1 and 1.0 mutations per Mb (about 250-2500 sSNVs per genome). SCcaller  
 951 standard uses a calling threshold of  $\alpha = 0.05$  while stringent calling uses  $\alpha = 0.01$ .  
 952

953



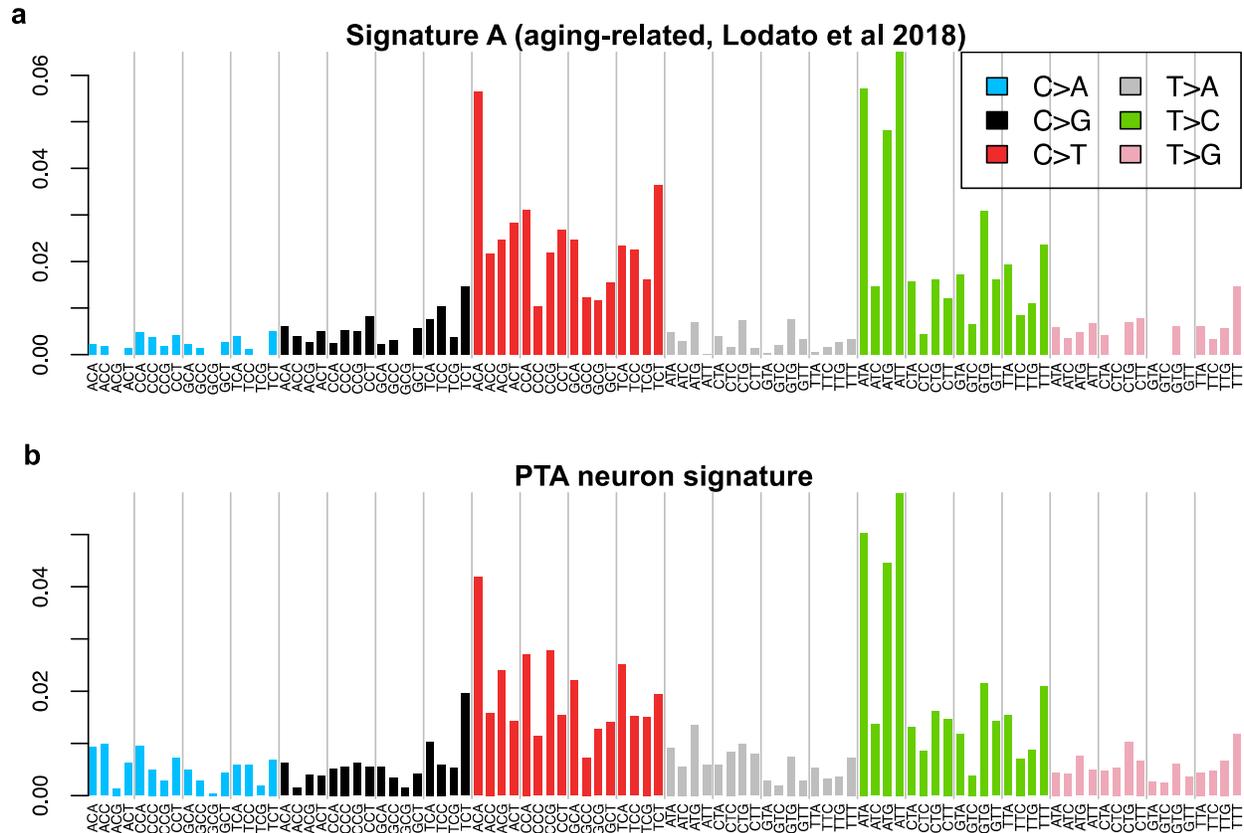
954

955

956 **Supplementary Figure 8: SCAN2 sensitivity on simulated indels.** a. Length distribution of  
 957 all simulated spike-in indels (black) and recovered indels (red). b. Spike-in indel sensitivity by  
 958 length. c. Sensitivity for indel detection stratified by ID83 indel class. Dotted outlines: sensitivity  
 959 before applying cross-subject filtration.

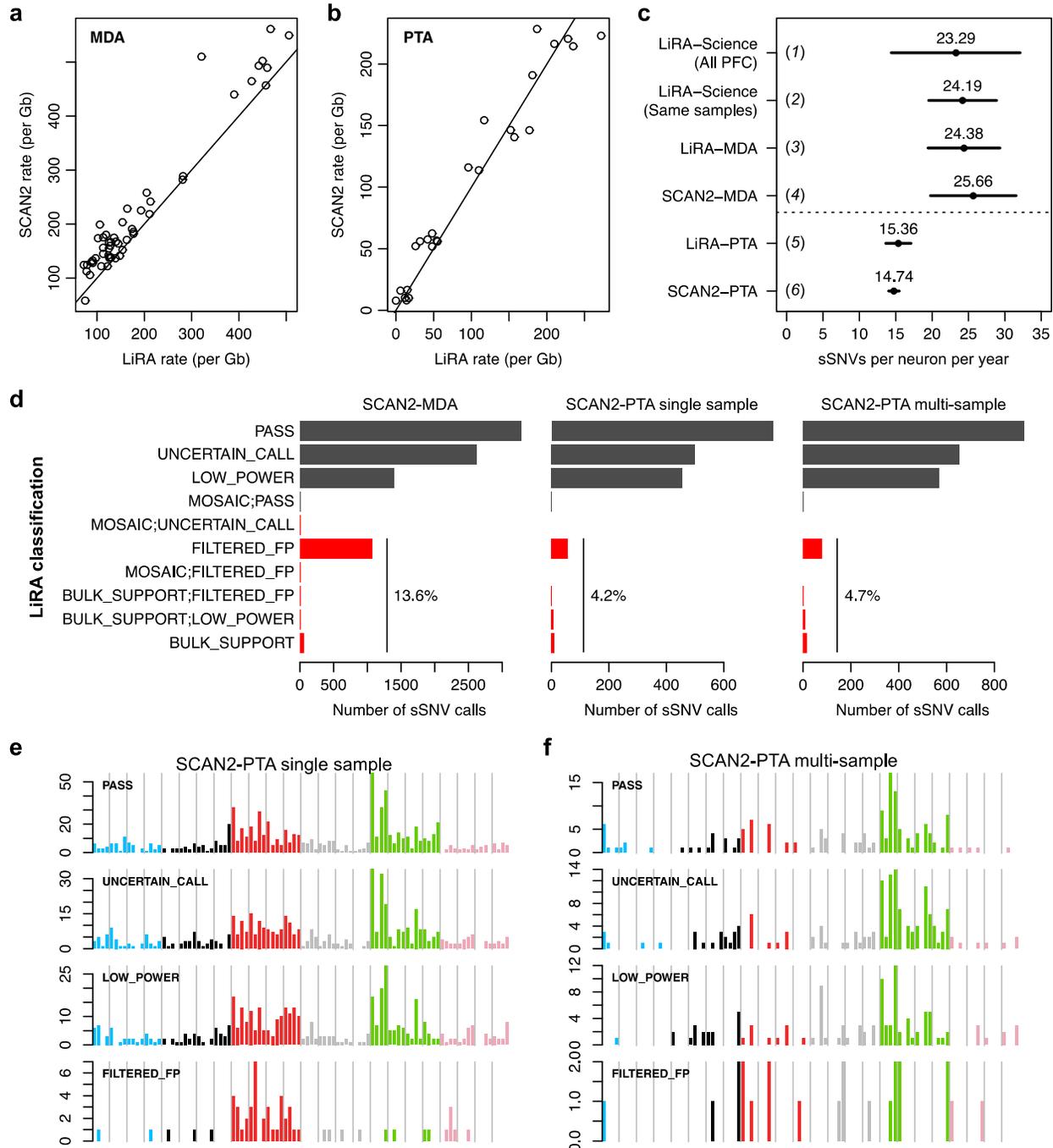
960

961



962  
963  
964  
965  
966  
967  
968  
969  
970  
971

**Supplementary Figure 9: PTA confirms the age-related sSNV signature in human neurons.** **a.** Aging-associated signature derived from MDA-amplified neurons (ref. 6). **b.** Mutation signature produced by single-sample SCAN2 on PTA-amplified human neurons. Multi-sample SCAN2 is not appropriate for mutation signature discovery because it is biased against mutations from signature components 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.

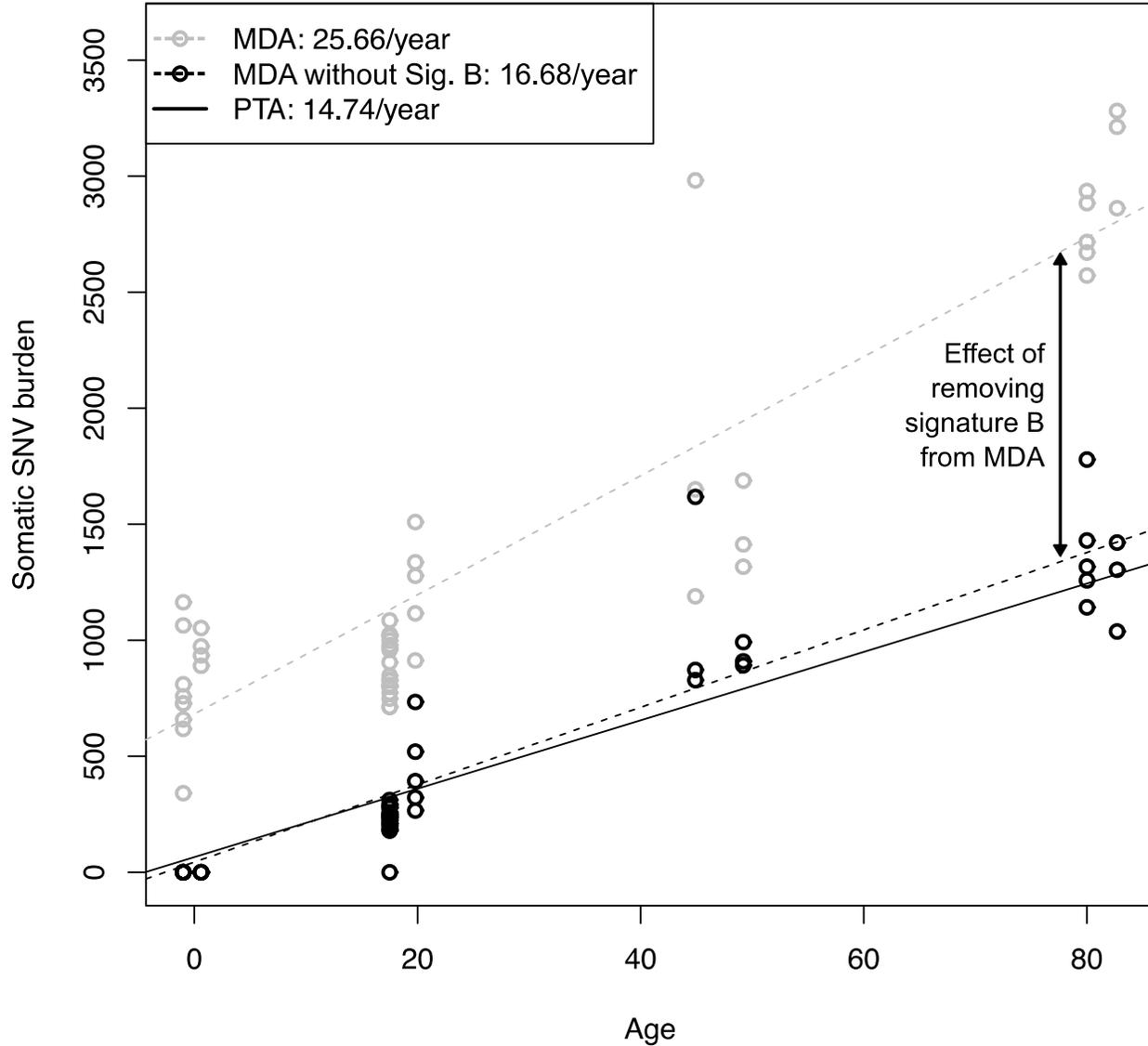


972  
973

974 **Supplementary Figure 10: Comparison of SCAN2 and LiRA on human neurons.** Single  
975 human neurons were previously analyzed by LiRA<sup>20</sup>, a specific but lower sensitivity approach for  
976 calling somatic SNVs. **a-b.** SCAN2 and LiRA extrapolations for the total (not called) sSNV  
977 burden per diploid Gb of human sequence from MDA- (**a**) and PTA-amplified (**b**) single neurons.  
978 Solid lines:  $y=x$ . **c.** Linear regression estimates for the number of sSNVs accumulated per  
979 neuron per year from several sources and analyses. Horizontal bars represent 95% C.I.s. (1)  
980 LiRA rates taken from ref. 6, which used a larger set of 91 MDA-amplified PFC neurons; (2)  
981 LiRA rates taken from ref. 6 using the same set of 51 MDA-amplified PFC neurons; (3) rerun of  
982 LiRA on 51 MDA-amplified neurons using the same input provided to SCAN2; (4) SCAN2 on 51

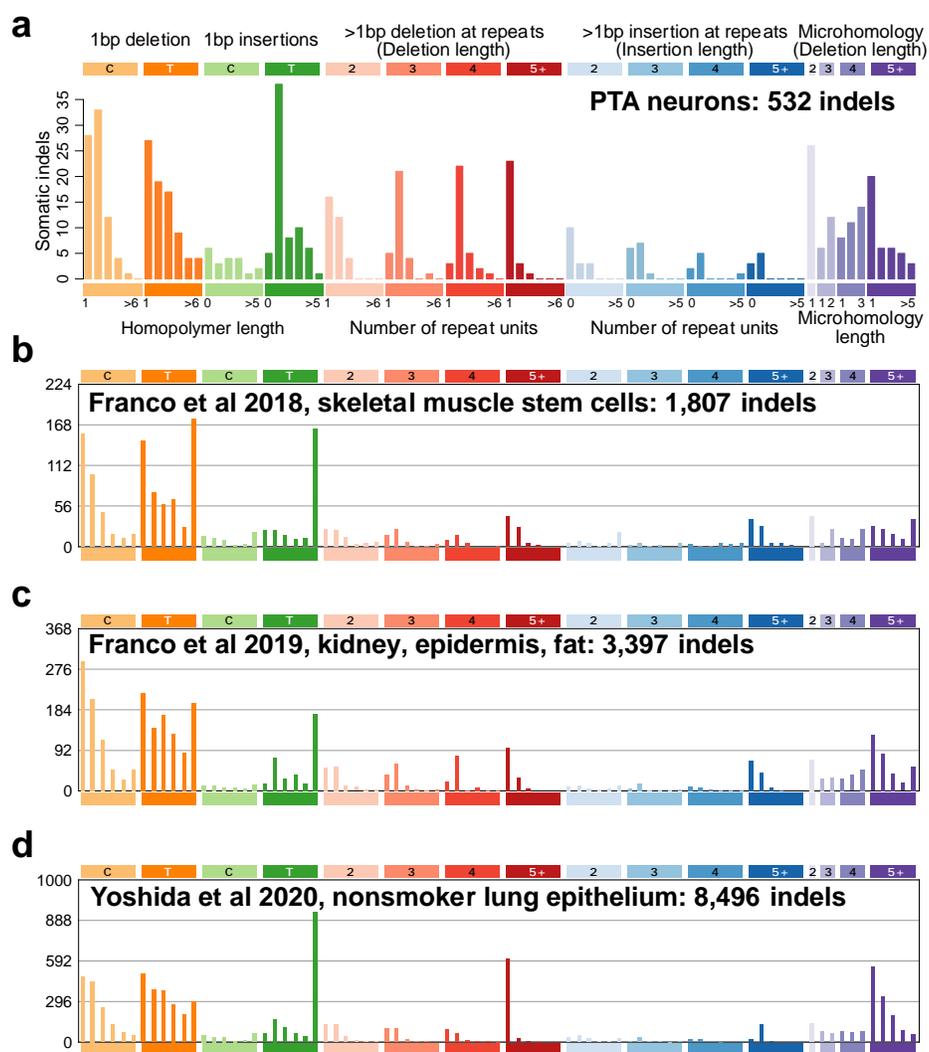
983 MDA-amplified neurons; (5) LiRA on 25 PTA-amplified neurons; (6) SCAN2 on 25 PTA-  
984 amplified neurons. **d.** LiRA classification of SCAN2 calls where reads linked to nearby germline  
985 heterozygous SNPs are available (black: likely true sSNVs, red: possible false positives). PASS  
986 is the highest quality LiRA class. UNCERTAIN and LOW\_POWER indicate lack of linking reads  
987 to make a confident call, but no evidence of artifactual status is detected. All other classes (red)  
988 are interpreted as false positives. Percentages show the fraction of all false positive classes  
989 among SCAN2 calls. **e-f.** Raw mutation spectra for single- (**e**) and multi-sample (**f**) SCAN2 calls  
990 stratified by LiRA classification. The similarities between PASS and the two lower quality  
991 UNCERTAIN\_CALL and LOW\_POWER classes suggest that the majority of  
992 UNCERTAIN\_CALL and LOW\_POWER SCAN2 calls are true mutations. Confident false  
993 positives (FILTERED\_FPs) possess a C>T dominated signature with lack of C>Ts at CpGs.

994  
995



996  
997  
998  
999  
1000  
1001  
1002  
1003

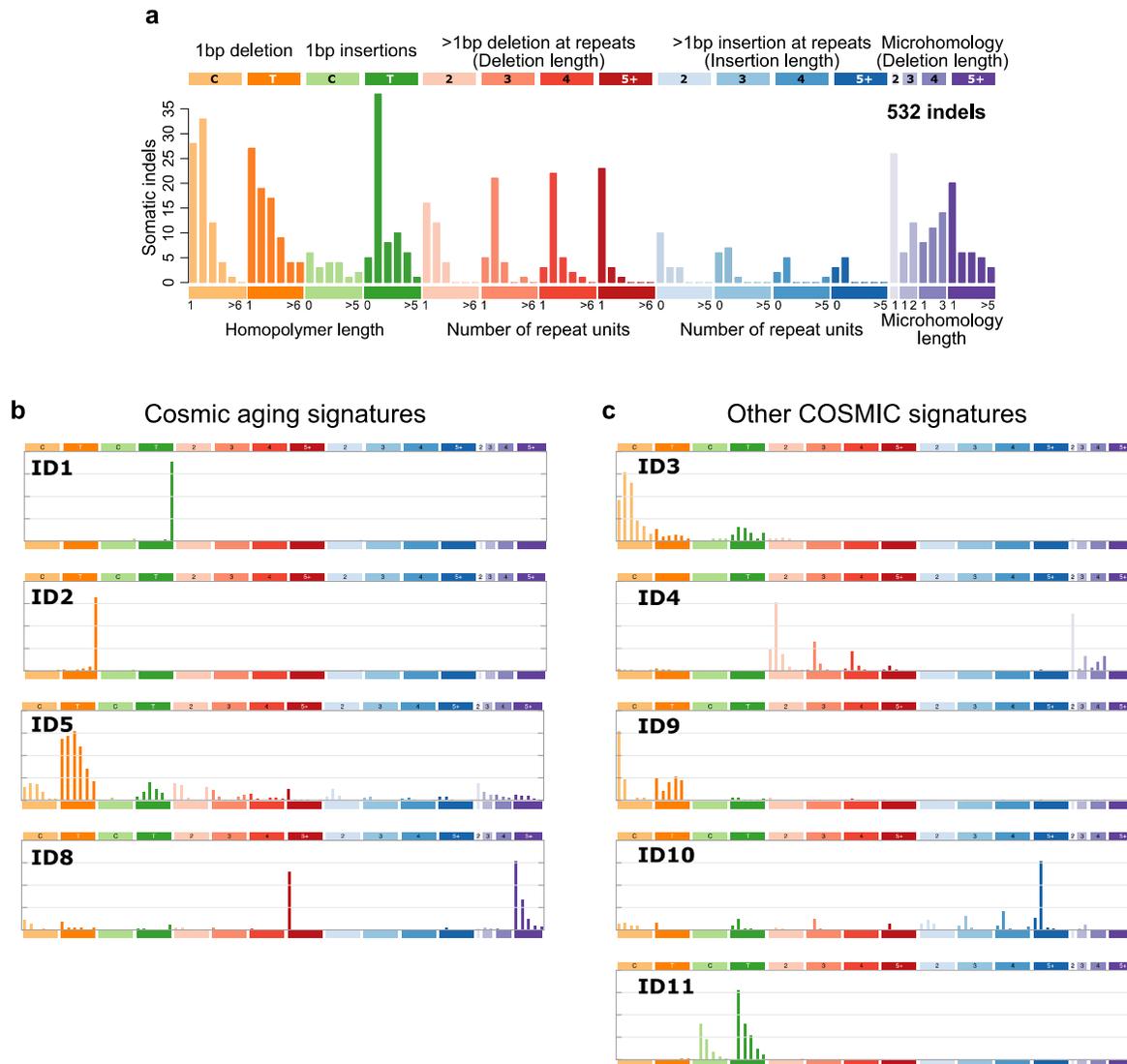
**Supplementary Figure 11: Removal of Signature B from MDA neurons closely matches PTA-derived mutation rates.** Total SCAN2-called somatic SNV mutation burdens from MDA neurons before Signature B removal (grey circles) and after Signature B removal (black circles). Trend lines: MDA accumulation rate (dotted grey), MDA accumulation rate after Signature B removal (dotted black), PTA accumulation rate (solid black).



1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012

**Supplementary Figure 12: Somatic indel signatures compiled from other publications. a.** 532 indels from PTA neurons from this study, same as **Figure 3f**. **b.** Clonally expanded single skeletal muscle stem cells. **c.** Clonally expanded single kidney, epidermis and fat cells. Excludes hypermutated kidney cells (designated KT2 in the original study). **d.** Clonally expanded bronchial epithelial cells from children and never-smokers.

1013  
1014



1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023

**Supplementary Figure 13: COSMIC indel signatures.** **a.** Spectrum of indels from PTA neurons (same as **Figure 3f**). **b.** COSMIC signatures with clock-like or age-associated annotations. **c.** Non-aging COSMIC signatures found in single neurons.

<b>Donor ID</b>	<b>Age</b>	<b>Sex</b>	<b>MDA</b>	<b>PTA</b>
<b>Infant</b>				
1278	0.4	M	9	3
5817	0.6	M	4	3
<b>Adolescent</b>				
1465	17.5	M	18	4
5559	19.8	F	5	3
<b>Adult</b>				
5087	44.9	M	4	3
936	49.2	F	3	3
<b>Aged</b>				
5657	82	M	5	3
5823	82.7	F	3	3

1024

1025

1026 **Supplementary Table 1: Individuals sequenced in this study.** Individuals from four age groups,  
1027 ranging from infants to the elderly, were analyzed in this study. MDA and PTA columns refer to  
1028 the number of PFC neurons amplified by each method and sequenced to high coverage.

1029

1030

Subject	Sample	Amp.	Age	Sex	Callable bp	MAPD
1278	1278_ct_p1E3	MDA	0.4	M	2377463116	0.582
1278	1278_ct_p1E6	MDA	0.4	M	2254957388	0.767
1278	1278_ct_p1G9	MDA	0.4	M	2310472262	0.717
1278	1278_ct_p2B9	MDA	0.4	M	2294524648	0.708
1278	1278_ct_p2C7	MDA	0.4	M	2351946883	0.727
1278	1278_ct_p2E4	MDA	0.4	M	2277857833	0.744
1278	1278_ct_p2E6	MDA	0.4	M	2315010769	0.73
1278	1278_ct_p2F5	MDA	0.4	M	2298927433	0.71
1278	1278_ct_p2G5	MDA	0.4	M	2285597264	0.722
1278	1278BA9-A	PTA	0.4	M	2559227873	0.188
1278	1278BA9-B	PTA	0.4	M	2564412679	0.187
1278	1278BA9-C	PTA	0.4	M	2570780791	0.186
1278	1278_heart_bulk	none	0.4	M	NA	
5817	5817_ct_p1H10	MDA	0.6	M	2218940766	0.827
5817	5817_ct_p1H2	MDA	0.6	M	2264476042	0.754
5817	5817_ct_p1H5	MDA	0.6	M	2280282603	0.753
5817	5817_ct_p2H6	MDA	0.6	M	2241619365	0.768
5817	5817PFC-A	PTA	0.6	M	2458549871	0.232
5817	5817PFC-B	PTA	0.6	M	2394321853	0.226
5817	5817PFC-C	PTA	0.6	M	2432653880	0.214
5817	5817_liver_bulk	none	0.6	M	NA	
1465	1465-cortex_1-neuron_MDA_12	MDA	17.5	M	2368914601	0.576
1465	1465-cortex_1-neuron_MDA_18	MDA	17.5	M	2328801902	0.569
1465	1465-cortex_1-neuron_MDA_20	MDA	17.5	M	2343394691	0.549
1465	1465-cortex_1-neuron_MDA_24	MDA	17.5	M	2250564147	0.63
1465	1465-cortex_1-neuron_MDA_25	MDA	17.5	M	2317117886	0.574
1465	1465-cortex_1-neuron_MDA_2_WGSb	MDA	17.5	M	2278772444	0.553
1465	1465-cortex_1-neuron_MDA_30	MDA	17.5	M	2278027817	0.607
1465	1465-cortex_1-neuron_MDA_39	MDA	17.5	M	2281108217	0.61
1465	1465-cortex_1-neuron_MDA_3_WGSb	MDA	17.5	M	2246826723	0.58
1465	1465-cortex_1-neuron_MDA_43	MDA	17.5	M	2311323225	0.545
1465	1465-cortex_1-neuron_MDA_46	MDA	17.5	M	2329270490	0.565
1465	1465-cortex_1-neuron_MDA_47	MDA	17.5	M	2276931799	0.57
1465	1465-cortex_1-neuron_MDA_5	MDA	17.5	M	2283876392	0.579
1465	1465-cortex_1-neuron_MDA_51_WGSb	MDA	17.5	M	2220441876	0.602
1465	1465-cortex_1-neuron_MDA_6_WGSb	MDA	17.5	M	2248579628	0.561
1465	1465-cortex_1-neuron_MDA_8	MDA	17.5	M	2319210026	0.628
1465	1465_ct_8p2h8	MDA	17.5	M	2396680438	0.548
1465	1465_ctx_p2g8	MDA	17.5	M	2346370110	0.58
1465	1465BA9-A	PTA	17.5	M	2502902455	0.207
1465	1465BA9-B	PTA	17.5	M	2379712047	0.28
1465	1465BA9-C	PTA	17.5	M	2490365389	0.232
1465	1465BA9-D	PTA	17.5	M	2385020412	0.272
1465	1465-cortex_BulkDNA_WGSb	none	17.5	M	NA	
5559	5559-pfc1C4	MDA	19.8	F	2335438836	0.696
5559	5559-pfc1C7	MDA	19.8	F	2219634045	0.819
5559	5559-pfc1E2	MDA	19.8	F	2243450288	0.861
5559	5559-pfc1H2	MDA	19.8	F	2177525787	0.815
5559	5559-pfc2A3	MDA	19.8	F	2380863121	0.701
5559	5559PFC-A	PTA	19.8	F	2506510681	0.21

5559	5559PFC-B	PTA	19.8	F	2474056125	0.206
5559	5559PFC-C	PTA	19.8	F	2532204421	0.193
5559	5559-bulk	none	19.8	F	NA	
5087	5087pfc-Lp1C5	MDA	44.9	M	1956709694	1.105
5087	5087pfc-Rp1G4	MDA	44.9	M	2045937656	1.027
5087	5087pfc-Rp3C5	MDA	44.9	M	851472780	1.758
5087	5087pfc-Rp3F4	MDA	44.9	M	2219472588	0.848
5087	5087PFC-A	PTA	44.9	M	2526638419	0.192
5087	5087PFC-B	PTA	44.9	M	2529648486	0.199
5087	5087PFC-C	PTA	44.9	M	2496175648	0.194
5087	5087-hrt-1b1	none	44.9	M	NA	
936	936_20141001-pfc-1cp1G11_20170221-WGS	MDA	49.2	F	2069054494	0.95
936	936_20141001-pfc-1cp1H9_20170221-WGS	MDA	49.2	F	2239568087	0.85
936	936_20141001-pfc-1cp2F6_20170221-WGS	MDA	49.2	F	1937342351	1.036
936	936PFC-A	PTA	49.2	F	2458178187	0.189
936	936PFC-B	PTA	49.2	F	2498078321	0.186
936	936PFC-C	PTA	49.2	F	2448162449	0.183
936	936-hrt-1b1_20170221-WGS	none	49.2	F	NA	
5657	5657-pfc1D2	MDA	82	M	1970196085	1.076
5657	5657-pfc1E11	MDA	82	M	2358110993	0.771
5657	5657-pfc2A6	MDA	82	M	2379723437	0.74
5657	5657-pfc2F1	MDA	82	M	2397069500	0.728
5657	5657-pfc2G9	MDA	82	M	2405157253	0.748
5657	5657PFC-A	PTA	82	M	2477582773	0.191
5657	5657PFC-B	PTA	82	M	2531288033	0.187
5657	5657PFC-C	PTA	82	M	2467010743	0.185
5657	5657-bulk	none	82	M	NA	
5823	5823_20160824-pfc-1cp1F11_20170221-WGS	MDA	82.7	F	2096166634	1.078
5823	5823_20160824-pfc-1cp2E1_20170221-WGS	MDA	82.7	F	1878096891	1.143
5823	5823_20160824-pfc-1cp2G5_20170221-WGS	MDA	82.7	F	1901575408	1.062
5823	5823PFC-A	PTA	82.7	F	2435263867	0.194
5823	5823PFC-B	PTA	82.7	F	2399384951	0.215
5823	5823PFC-C	PTA	82.7	F	2494418160	0.19
5823	5823-tempmusc-1b1_20170221-WGS	none	82.7	F	NA	

1031

1032 **Supplementary Table 2: Samples analyzed in this study.** List of all samples used in this study.

1033 For single cell samples, the method of genome amplification is listed (MDA or PTA); samples  
 1034 with amplification “none” are bulk controls. Callable bp indicates the number of base pairs in  
 1035 the human genome which passed basic depth criteria for analysis (>5 in the single cell, >10 in  
 1036 the matched bulk).

1037

1038