

This is a repository copy of *Somatic mutation landscapes at single-molecule resolution*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/188385/>

Version: Accepted Version

Article:

Abascal, Federico, Harvey, Luke M R, Mitchell, Emily et al. (27 more authors) (2021) Somatic mutation landscapes at single-molecule resolution. *Nature*. 405–410. ISSN 0028-0836

<https://doi.org/10.1038/s41586-021-03477-4>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

51 rates and signatures to a variety of mitotically-active tissues. Together these results suggest
52 that mutational processes independent of cell division are important contributors to adult
53 somatic mutagenesis. We anticipate that the ability to reliably detect mutations in single
54 molecules of DNA could transform our understanding of mutagenesis in vivo and in vitro, and
55 enable somatic mutation studies in large-scale cohorts.

57 **Introduction**

58
59 Somatic mutations occur in our cells as we age, driving cancer development and potentially
60 contributing to ageing and other diseases. Despite their importance, their study remains
61 challenging due to technical limitations. Because any given somatic mutation in a normal tissue
62 is typically present in a small group of cells or even in a single cell, detecting them requires
63 special approaches, such as ultra-deep sequencing of small biopsies³⁻⁵, laser microdissection⁶⁻
64 ⁸, isolation of single-cells followed by in vitro expansion into organoids or colonies⁹⁻¹¹, or
65 single-cell sequencing¹²⁻¹⁴. While these technologies are changing our understanding of
66 somatic mutagenesis, the error rate of single-cell sequencing remains high¹⁵, and other
67 approaches are typically limited to mitotically-active cell types.

68
69 As a result of technical limitations, the rates and patterns of somatic mutation across most
70 human cell types remain underexplored. This is especially the case for non-dividing cells,
71 including the differentiated cells that make up the bulk of mitotically-active tissues and are
72 responsible for tissue function, as well as post-mitotic tissues, such as cortical neurons or
73 cardiac muscle, which are of particular interest in human ageing, neurodegeneration and
74 cardiovascular disease. Post-mitotic tissues can also shed light on the contribution of cell
75 division and DNA replication to somatic mutation in human tissues. To address these
76 questions, here we develop a new sequencing protocol that reliably detects mutations in single
77 molecules of DNA from populations of cells, enabling the study of somatic mutation in any
78 tissue or cell population.

80 **NanoSeq achieves error rates two orders of magnitude below somatic mutation rates**

81
82 The fundamental limitation of standard sequencing methods for the study of genetically
83 heterogenous samples is the need to detect the same mutation in multiple cells to distinguish
84 genuine mutations from sequencing errors, a consequence of their error rates being above 10^{-3}
85 errors per base pair (bp)¹⁶. Several protocols have been developed to increase the accuracy of
86 standard sequencing methods by tagging individual molecules of DNA with unique molecular
87 barcodes and reading the same molecule multiple times, reducing error rates by single-
88 molecule consensus sequencing¹⁶. The most accurate approaches are based on duplex
89 consensus sequencing^{17,18}, which rely on sequencing copies of both strands of a DNA molecule
90 to remove sequencing errors (present in individual reads) and PCR errors (present in copies of
91 one of the two strands) (**Fig 1a**).

92
93 Duplex sequencing has a theoretical error rate $<10^{-9}$ errors/bp, which is the probability of two
94 early and complementary PCR errors in both strands¹⁶. Given that this theoretical limit is lower
95 than the typical mutational load of human tissues, it raises the possibility of quantifying somatic
96 mutation rates in any cell type, independently of its clonal architecture. This is the rationale of
97 BotSeqS, a whole-genome duplex sequencing protocol¹⁹. In practice, however, mapping errors
98 and the accidental copying of errors between strands during library preparation violate the
99 independence of both strands and limit the accuracy of duplex sequencing^{19,20}. The actual error
100 rates of duplex sequencing protocols have remained difficult to measure¹⁶, but some protocols

101 report error rates above 10^{-7} errors/bp²⁰, translating into hundreds to thousands of errors per
102 diploid genome.

103

104 A difficulty in measuring the error rate of existing duplex sequencing protocols has been the
105 lack of control samples with known mutation rates. To evaluate the performance of the existing
106 BotSeqS protocol (**Fig 1a**), we first analysed a sample of granulocytes from a 59-year-old
107 donor from whom 110 single-cell derived blood colonies had been whole-genome sequenced²¹
108 (**Supplementary Table 1,2**). We found that the estimates of mutation burden per diploid
109 genome from BotSeqS were two-fold higher than those from the colonies (**Fig 1b**), and that
110 the substitution profiles were dissimilar (cosine similarity of 0.71; **Fig 1c**), with increased C>A
111 and C>G substitution rates. Analysing the distribution of substitutions across the reads revealed
112 a large excess of G>T/C substitutions near the 5' ends of DNA fragments, and an imbalance
113 over C>A/G substitutions that affected the entire read length (**Fig. 1d** and **Extended Data**
114 **Figures 1 and 2**). These substitution imbalances are incompatible with real mutations and
115 reflect errors introduced during library preparation²² (**Methods, Supplementary Note 1**). We
116 confirmed that the same imbalances, together with an additional C>T asymmetry, were present
117 in the original BotSeqS publication¹⁹ (**Fig 1d**). Extensive trimming of read ends only partially
118 alleviated these errors (**Extended Data Fig 2**). Based on these results, we estimate that BotSeqS
119 introduced ~1,500 errors per diploid genome in our samples, equivalent to an error rate $\sim 2.6 \times$
120 10^{-7} errors/bp.

121

122 DNA damage in one strand can be fixed as an apparent mutation in both DNA strands during
123 end repair, violating the error-correction mechanism of duplex sequencing (**Fig 1e, Extended**
124 **Data Fig 1c-d**). To solve this, we developed NanoSeq, a protocol that prevents copying errors
125 between strands by avoiding end repair and by blocking nick extension. First, we replaced
126 sonication and end repair with restriction enzyme fragmentation (**Fig 1e**). We chose HpyCH4V
127 based on in silico estimations of achievable genomic coverage (**Methods; Supplementary**
128 **Table 3; Supplementary Note 2**). Although restriction enzymes provide partial coverage of
129 the genome (29% using HpyCH4V), the fraction covered is sufficiently random to accurately
130 estimate mutation rates and signatures, and they enable the generation of NanoSeq libraries
131 from as little as 1 ng of DNA (**Methods**). Alternatively, we show that sonication followed by
132 exonuclease blunting can be used for applications requiring whole-genome coverage
133 (**Methods, Supplementary Note 3, Extended Data Fig 3**). Second, we introduced dideoxy
134 non-A nucleotides (ddBTPs) during A-tailing, to avoid errors from nick extension (**Fig 1e;**
135 **Methods; Extended Data Fig 1e; Supplementary Note 4**). Adapters with sufficiently diverse
136 random barcodes were used to tag PCR duplicate families (**Supplementary Note 5**). As it is
137 standard in somatic mutation calling, a polyclonal matched normal sample is used alongside
138 NanoSeq to distinguish germline and somatic mutations (**Methods**).

139

140 Duplex sequencing and BotSeqS often suffer from low efficiency due to suboptimal recovery
141 of reads from both original strands. We show that mathematical modelling of family sizes and
142 qPCR quantification of the library can be used to maximise the duplex coverage independently
143 of the amount of input DNA (**Methods, Extended Data Fig 4a-d**). A robust bioinformatic
144 pipeline was developed to avoid false positive mutation calls from mapping errors or low-level
145 DNA contamination (**Extended Data Fig 4e,f; Methods; Supplementary Note 6**), and to
146 distinguish germline from somatic mutations.

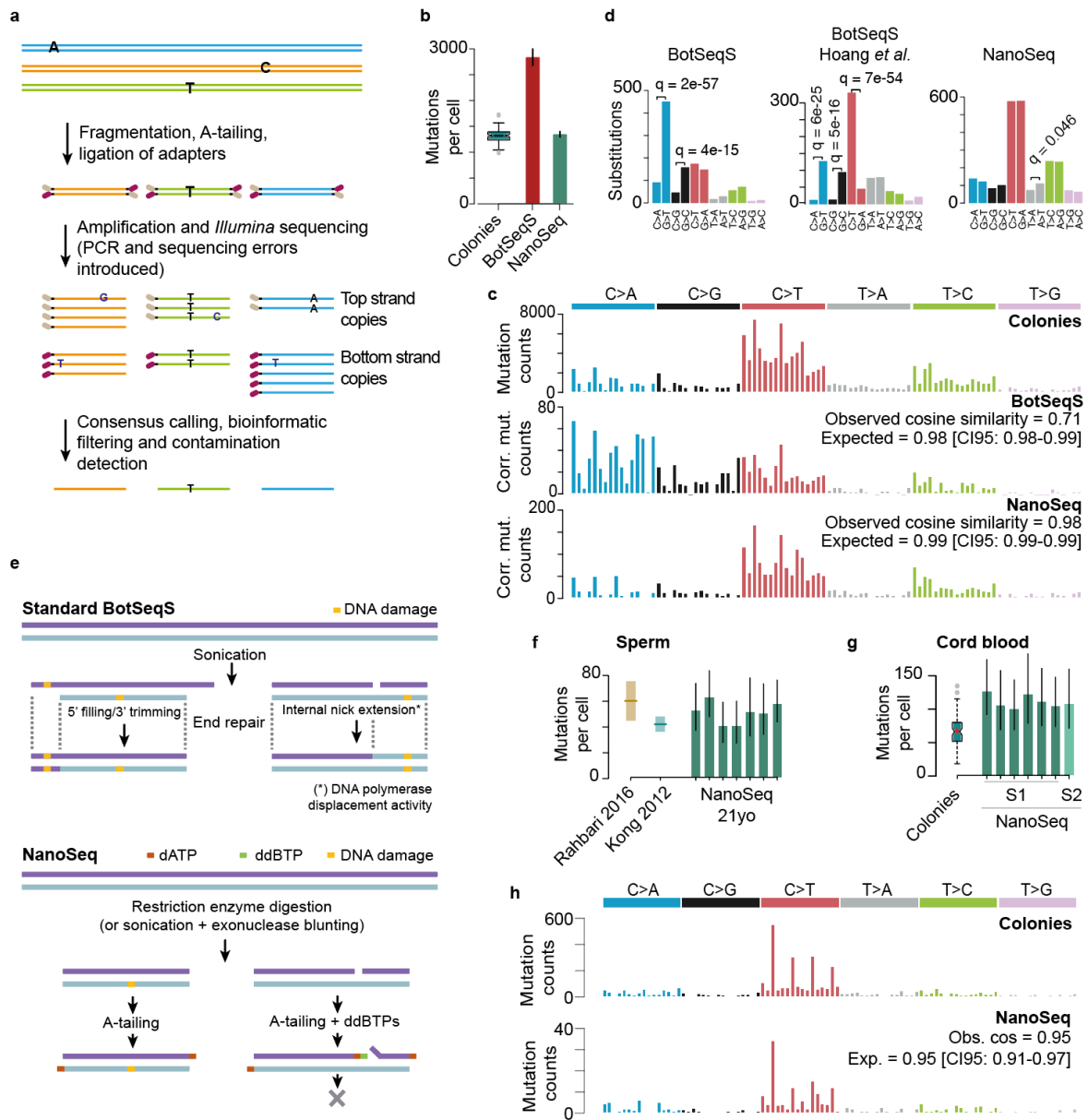
147

148 Applying the NanoSeq protocol to the same sample of granulocytes from the 59-year-old donor
149 (**Supplementary Table 1,2**), yielded nearly-identical burden estimates and substitution
150 profiles to the colonies (cosine similarity of 0.98) (**Fig 1c; Methods; Supplementary Note 7;**

151 **Extended Data Fig 5a,b**). We detected no evidence of substitution imbalances except for a
152 slight enrichment of A>T over T>A, which we have not seen in subsequent libraries (**Fig 1d**).
153 To measure the error rate of NanoSeq we then applied it to samples with low mutation burdens:
154 a sperm sample from a 21-year-old donor and cord blood granulocytes from two neonates.
155 Seven replicates of the sperm sample yielded low mutation burdens, with ~52 mutations per
156 haploid sperm cell (1.8×10^{-8} mutations/bp or ~2.5 mutations/year/cell), consistent with current
157 estimates of the mutation rate in the paternal germline from trio studies^{23,24} (**Fig 1f**). NanoSeq
158 estimates from cord blood granulocytes were compared to 100 single-cell derived cord blood
159 colonies from two different donors. Corrected NanoSeq estimates (**Methods**) were higher than
160 those from blood colonies (109 vs 66 mutations per cell; 95% Poisson confidence intervals 95-
161 125; **Fig 1g**). This difference could be due to NanoSeq errors, higher burden in granulocytes
162 than stem-cell-derived colonies, or both. Consistent with most mutations detected by NanoSeq
163 being genuine, comparison of both mutational spectra did not detect significant differences
164 between them (**Fig 1h, Methods**).

165
166 Together, the sperm and cord blood data indicate that the error rate of NanoSeq is considerably
167 lower than 5×10^{-9} errors/bp (<30 errors per diploid genome), two orders of magnitude lower
168 than the BotSeqS error rate and the somatic mutation load of most human tissues studied to
169 date. Analysis of insertions and deletions (indels) in cord blood similarly confirms that the
170 NanoSeq indel error rate is $<3 \times 10^{-9}$ errors/bp (**Methods; Extended Data Fig 5c;**
171 **Supplementary Note 8**).

172
173 To our knowledge, these are the lowest confirmed error rates of any DNA sequencing protocol.
174 These error rates open the door to the accurate study of somatic mutations in any tissue type,
175 independent of clonality. We take advantage of this unprecedented ability to reliably study
176 non-dividing cells across four tissues, addressing two elusive questions in the field of somatic
177 mutagenesis: the difference in mutation rates between stem cells and terminally-differentiated
178 cells in mitotically-active tissues, and the rates and patterns of mutation in post-mitotic tissues.
179
180



182
 183 **Figure 1 | Standard BotSeqS and NanoSeq sequencing protocols.** **a**, Fundamentals of duplex sequencing
 184 protocols. **b**, Mutation burden estimates in granulocytes using BotSeqS and NanoSeq, compared to standard
 185 results with single-cell derived blood colonies. Box plot show the interquartile range, median and 95%
 186 confidence interval for the median. BotSeqS and NanoSeq bars show 95% Poisson confidence intervals. **c**,
 187 Comparison of BotSeqS and NanoSeq granulocyte substitution profiles with blood colonies data (the
 188 calculation of expected cosine similarities is explained in the **Methods** section). The same filtering
 189 approaches were used for both BotSeqS and NanoSeq. **d**, Substitution imbalances are present in standard
 190 BotSeqS protocols but absent from NanoSeq (**Extended Data Figs 1a,b** and **2**). Imbalances were tested with
 191 a binomial test assuming p of 0.5 and p-values were corrected with Benjamini and Hochberg's FDR method.
 192 **e**, Standard BotSeqS (top) and the new NanoSeq approach (bottom) for genome fragmentation and library
 193 preparation. **f**, NanoSeq mutation burden estimates for seven sperm samples from a 21-year-old donor
 194 compared to reported estimates of mutation burden in sperm, showing 95% Poisson confidence intervals. **g**,
 195 NanoSeq mutation burden estimates for cord blood granulocytes compared to single-cell derived cord blood
 196 colonies, showing 95% Poisson confidence intervals; Box plot show the interquartile range, median and 95%
 197 confidence interval for the median, with the mean and its 95% confidence interval shown in red. **h**,
 198 Comparison between cord blood colonies and granulocyte substitution profiles.
 199

200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249

Similar mutation burden in stem and differentiated cells, in blood and colon

Most of our knowledge of mutagenesis in normal tissues is restricted to stem or proliferating cells. Since stem cells are believed to be genetically more protected than differentiated cells²⁵, differentiated cells could conceivably have higher mutational loads and undescribed mutational signatures¹⁴.

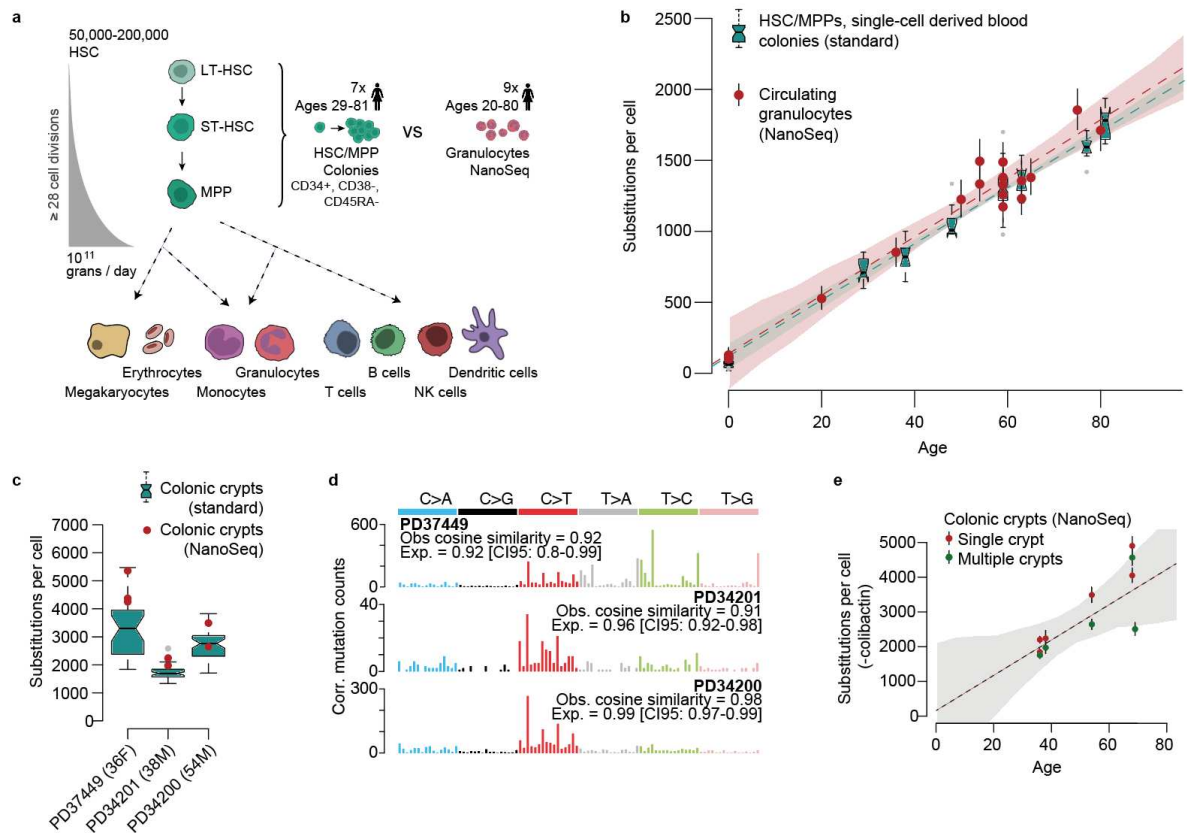
We first addressed this question in the haematopoietic system, comparing the mutational landscape of mature granulocytes to that of haematopoietic stem cell and multipotent progenitor cells (HSC/MPPs) (**Methods**). The haematopoietic system is organised hierarchically, with a heterogeneous pool of slow-cycling stem cells at the top of the hierarchy sustaining the production of large numbers of differentiated cells through the extensive proliferation of intermediate progenitor cells (**Fig 2a**). Given the number of divisions separating slow-cycling stem cells and granulocytes, a considerably higher mutation burden in granulocytes as well as mutational signatures associated with proliferation may be expected. We used NanoSeq to sequence 18 samples of granulocytes from 9 healthy donors, ranging from 20 to 80 years of age (**Supplementary Table 1,2**). We compared these data to standard whole-genome sequencing of 60 single-cell derived HSC/MPPs colonies from 6 donors (**Extended Data Fig 6a; Supplementary Table 1,2**) and published data from 110 colonies from one donor²¹ (**Methods**).

These data revealed that terminally-differentiated granulocytes have remarkably similar mutation burdens and mutational signatures to HSC/MPPs (**Fig 2a**). Linear mixed-effect regression reveals indistinguishable slopes for HSC/MPPs colonies and granulocytes ($P=0.90$), with a combined estimate of ~ 19.8 mutations/year (CI95% 18.3-21.4, **Methods**). This slope, which reflects the accumulation of somatic mutations with age, provides an estimate of the mutation rate in the stem cells responsible for long-term maintenance of the haematopoietic system. Measured as the difference between intercepts, the excess of mutations in granulocytes over HSC/MPPs colonies is estimated to be ~ 57.7 mutations and not significantly different from zero (CI95%: -13.1-121.1, $P=0.12$, **Methods**).

The similarity in mutation burden and mutational signatures between granulocytes and HSC/MPPs is surprising given that HSC/MPPs are expected to have undergone many fewer cell divisions on average. HSCs are believed to divide around once a year and our conservative estimates suggest that at least an average of 28 additional divisions must separate stem cells from differentiated cells to explain the production of $\sim 10^{14}$ mature cells per year (**Fig 2a; Supplementary Note 9**). The observation that a considerable increase in cell divisions does not cause a proportional increase in mutation burden suggests that replication errors are only responsible for a minority of the mutations that occur in haematopoietic stem cells (**Supplementary Note 9**).

A caveat for the comparison between HSC/MPPs colonies and granulocytes is that HSC/MPPs are a heterogeneous population and estimates of mutation burden from colonies successfully grown in vitro may not reflect the mutation rate of the more quiescent stem cells responsible for long-term maintenance of the haematopoietic system. However, a similar conclusion can be drawn from the regression data on granulocytes alone, without comparison to the HSC/MPPs colonies. The strong linear relationship with age and the small intercept for granulocytes alone (157.4 mutations, CI95%: -106.4-423.5, compared to the slope of ~ 19.8 mutations/year) suggests that the majority of the mutations observed in adult granulocytes

250 accumulated in stem cells responsible for long-term maintenance, and that only a small
 251 minority of mutations are accrued during transient proliferation and terminal differentiation
 252 (Supplementary Note 9).
 253



254
 255 **Figure 2 | Mutation analyses of differentiated cells.** **a**, Schematic representation of the hematopoietic
 256 lineage showing which cell types and donors were analysed. **b**, Substitutions per cell for donors of different
 257 ages, comparing estimates from NanoSeq granulocytes (red) to standard sequencing of single-cell derived
 258 blood colonies (dark cyan boxplots); boxplots and confidence intervals as in Fig 1b; red and dark cyan
 259 dashed lines are linear mixed regression models; linear mixed model 95% confidence intervals for NanoSeq
 260 data calculated through parametric bootstrapping. For granulocytes, the intercept is 137.6 [CI95% -117.6-
 261 413.2] and the slope 20.6 [15.8-25.2]. For blood colonies, the intercept is 120.4 [27.9-218.5] and the slope
 262 19.8 [18.2-21.4]. **c**, Comparison between standard methods and NanoSeq burden estimates for colonic crypts
 263 from three donors. **d**, Substitution profiles for colonic crypts from the three donors and cosine similarities to
 264 profiles obtained with standard methods. **e**, Accumulation of substitutions throughout life in colonic crypts
 265 from 5 donors, excluding substitutions attributed to the episodic colibactin signature; confidence intervals
 266 as in panel b. Intercept of 156.9 [-1776.8-2117.7] and slope of 50.9 [9.8-91.1] (54.1 [43.0-64.9] without
 267 intercept).
 268

269 To extend the comparison of stem cells and differentiated cells to another tissue with a well-
 270 understood stem cell organisation, we studied colonic epithelium. Estimates of the somatic
 271 mutation rate in colonic stem cells are available from whole-genome sequencing of clonal
 272 organoids derived from Lgr5+ cells¹⁰ and from sequencing single colonic crypts⁶. Genome
 273 sequencing of whole crypts can be used to estimate the somatic mutation rate of colonic stem
 274 cells, as colonic crypts are clonally derived from a single stem cell. However, the process of
 275 reaching clonality through genetic drift in the population of stem cells within a crypt is
 276 estimated to take several years in humans²⁶, which could lead to an underestimation of mutation
 277 burdens using single-crypt sequencing.
 278

279 For three previously-studied donors we compared standard whole-genome sequencing of laser-
280 microdissected colonic crypts⁶ to NanoSeq data from single crypts or groups of crypts. This
281 revealed similar estimates of mutation burden, despite the lag to clonality in standard
282 sequencing (**Fig 2c**). Mutational burden and signatures from differentiated cells in colonic
283 epithelium were overall consistent with those found by previous studies on colonic stem cells,
284 with a dominance of SBS1, SBS5 and, in some donors, a colibactin signature²⁷ (**Fig 2d,e**).

285

286 Overall, NanoSeq data on granulocytes and colonic epithelium yielded similar estimates of
287 mutation burden and mutational signatures to their corresponding stem cells. While larger
288 studies will be needed to identify subtler differences in mutation rates between stem cells and
289 differentiated cells in granulocytes and colon, and to address this question in other cell types,
290 these results provide an early view into the somatic mutation landscape of two differentiated
291 cell types.

292

293 **Lifelong mutagenesis in post-mitotic neurons and polyclonal smooth muscle**

294

295 Cortical neurons are a prime example of a post-mitotic tissue. This makes them both a key cell
296 type to study somatic mutagenesis in the absence of cell division, and also inaccessible to
297 traditional sequencing methods. Single-cell sequencing has provided insights into somatic
298 mutation in neurons^{12,13}, although it remains unclear to what extent amplification artefacts
299 affected these results. Despite the technical challenges impeding progress, somatic mutation in
300 healthy neurons and in neurodegeneration has attracted considerable interest^{1,13,28,29}.

301

302 We applied NanoSeq to frontal cortex neurons from 8 healthy donors and 9 Alzheimer's disease
303 (AD) patients (**Supplementary Table 1**), using nuclei sorting with the *NeuN* neuronal marker
304 (**Methods; Extended Data Fig 7a**). These data revealed a tight linear accumulation of 20.0
305 substitutions (linear regression, CI95%:19.1-20.9) and 3.1 indels (CI95%:2.9-3.3) per year,
306 approximately constant throughout life (**Fig 3a,b**). This confirms that mutations accumulate in
307 a clock-like fashion in cortical neurons, in the absence of cell division, consistent with
308 observations from single-cell sequencing¹³.

309

310 These data shed new light on previously published single-neuron sequencing results. A study
311 using SNP-phased error-corrected single-cell sequencing reported three dominant signatures
312 in neurons, one that increased linearly with age and two that did not¹³. The spectrum found by
313 NanoSeq, the burden per genome and the mutation rate per year closely resemble the age-
314 associated signature in that study (cosine similarity 0.96; **Extended Data Fig 7b,c**). The other
315 two mutational signatures, responsible for around 72% of all mutations reported in the study
316 and highly variable across single-cell libraries (**Extended Data Fig 7d**), appear exclusively in
317 single-cell data and seem more consistent with amplification errors or transient DNA damage.
318 Consistent with this hypothesis, the dominant signature in single-neuron data closely resembles
319 a single-cell-specific signature reported in vitro¹⁵ (cosine similarity 0.97, **Extended Data Fig**
320 **7b**).

321

322 To better understand the mutational processes active in neurons in the absence of cell division,
323 we carried out signature decomposition on NanoSeq data from neurons together with data from
324 granulocytes, colonic crypts and smooth muscle (described below). Three signatures were
325 extracted (**Fig 3e**): signatures A and C imperfectly resembled SBS5 (cosine similarity 0.80)
326 and SBS16 (0.78), respectively, while signature B closely matched SBS1 (C>T changes at CpG
327 dinucleotides, cosine similarity 0.96). It is conceivable that SBS5, which appears to be a
328 ubiquitous signature in normal tissues and cancer genomes³⁰, reflects a collection of co-

329 occurring processes, rather than a single mutational process, leading to some differences across
330 tissues. The observation in post-mitotic neurons of signatures resembling SBS5 and SBS16
331 suggests that these common processes, whose aetiologies remain poorly understood, can occur
332 independently of cell division.

333
334 The substitution and indel spectra from neurons (**Fig 3c,d**) showed some differences with those
335 from granulocytes (**Fig 1c**) and smooth muscle (**Fig 3l,m**). T>C substitutions are more frequent
336 in neurons, especially at ApT dinucleotides (**Fig 3c**), and, together with C>G and C>T, show
337 strong transcriptional strand biases (**Extended Data Fig 8**). Interestingly, signature B (SBS1),
338 which is often assumed to be linked to cell division, accumulates at a low rate with age in
339 neurons (1.8 substitutions per year, linear regression CI95% 0.23-3.3, $P = 0.03$; **Extended**
340 **Data Fig 7e**). The presence of C>T mutations at CpG sites in neurons is better appreciated
341 normalising the rates by the trinucleotide frequency in the genome (**Extended Data Fig 9a,b**),
342 and implies that C>T mutations caused by 5-methylcytosine deamination can be fixed in both
343 DNA strands without cell division. In contrast to other somatic tissues, in neurons we did not
344 find a clear association between expression levels and substitution rates across genes (**Fig 3f**)
345 and the enrichment of mutations in heterochromatin was weaker (**Fig 3g**). Comparison of the
346 mutational spectra between active and inactive chromatin regions revealed different
347 contribution of the three mutational signatures across tissues (**Extended Data Fig 8a**).

348
349 Indel analysis revealed a higher relative frequency of indels in neurons than in other tissues,
350 caused by an unusual signature characterised by indels longer than 1bp (**Fig 3d,m**; **Extended**
351 **Data Fig 9c**). This indel signature and its association with highly expressed genes has some
352 resemblance to a little-understood mutational process recently described in cancer genomes³¹
353 (**Extended Data Fig 9d**).

354
355 Although the difference is small, AD donors showed a slightly lower substitution rate than
356 healthy donors (linear regression, 19.1 (CI95%:18.1-20.0) vs 21.6 (CI95% 20.5-22.7)
357 substitutions/year, $P = 0.006$). This difference was significant for signatures A and B but not
358 C ($p_A = 0.02$; $p_B = 0.03$; $p_C = 0.55$; **Fig 3i**; **Extended Data Fig 7e**). The difference in
359 mutation burden between controls and AD donors could merely reflect differences in the
360 patient cohorts or be related to the pathogenesis of the disease, for example due to differences
361 in metabolism or variable death rates across subpopulations of neurons in AD. Studies with
362 larger cohorts will be required to validate and explain this observation.

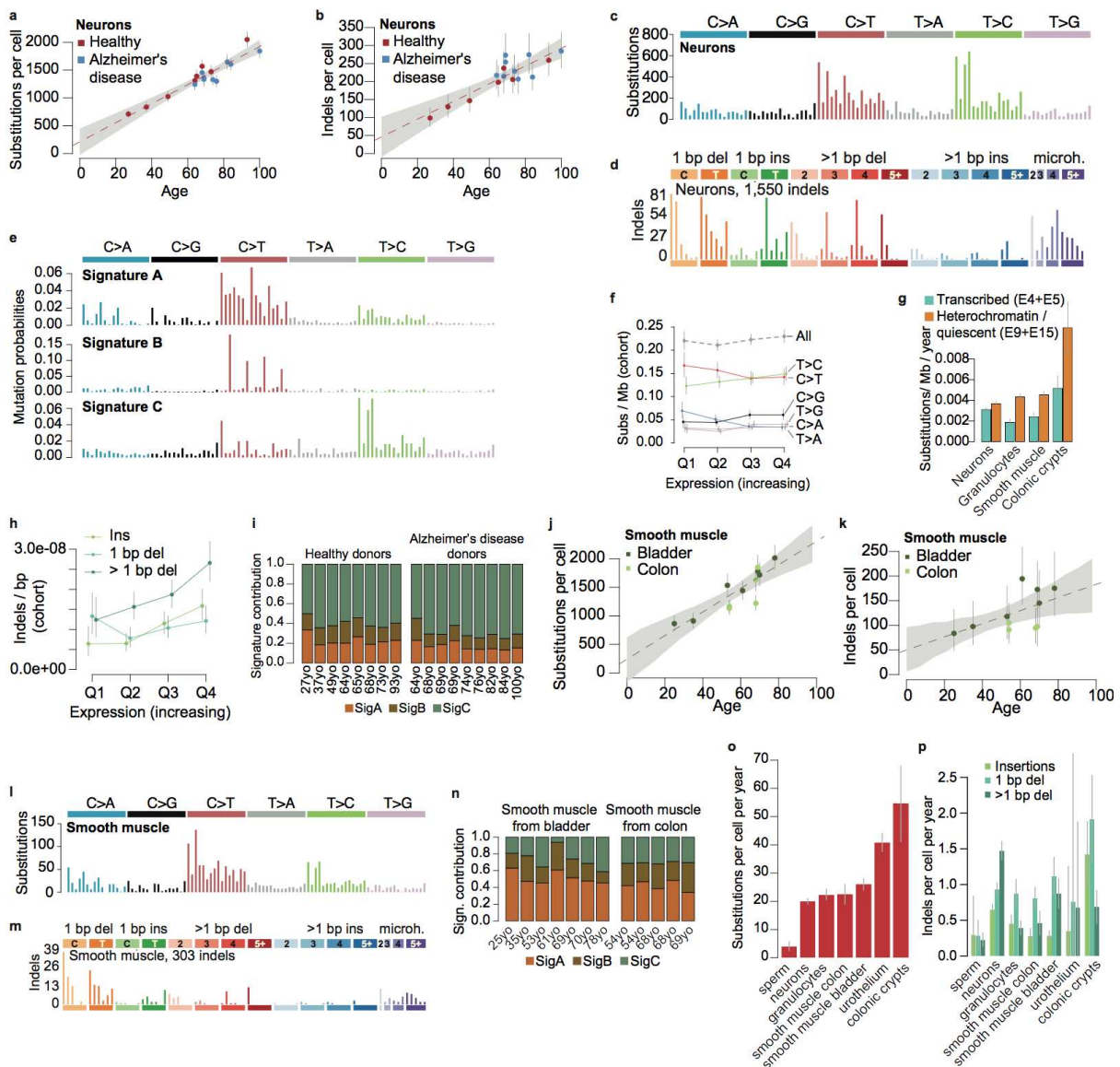
363
364 To extend these analyses to another tissue not amenable to standard sequencing methods, we
365 studied smooth muscle. Visceral smooth muscle cells are believed to divide infrequently in
366 normal conditions³². Using laser microdissection, we collected samples of smooth muscle from
367 10 donors and from two different organs, bladder and colon (**Supplementary Table 1,2**;
368 **Extended Data Fig 6b, 10a**). As expected for a polyclonal tissue, standard whole-genome
369 sequencing detected few mutations and at low allele frequencies in these samples (**Extended**
370 **Data Fig 10b,c, Methods**). In contrast, NanoSeq revealed that the substitution and indel
371 burdens increase linearly with age, with ~24.7 substitutions per year per diploid genome
372 (CI95%:22.5-27.0) and ~2.1 indels per year (95%:1.7-2.5) (**Fig 3j,k**). Despite their different
373 anatomical origin, smooth muscle cells from the bladder and colon walls showed relatively
374 similar mutation rates (mixed-effects linear regression, $P = 0.6$ for substitutions, $P = 0.04$ for
375 indels).

376
377 The mutation spectrum of smooth muscle partially resembled that of granulocytes (**Fig 3l,m**,
378 **Fig 1c**). All three signatures (A-C) accumulated linearly with age in smooth muscle (**Extended**

379 **Data Fig 7f)**, with similar contributions in smooth muscle from bladder and colon and across
 380 donors (**Fig 3n**). The smooth muscle spectra also resemble that of skeletal muscle satellite cells,
 381 studied by in vitro expansion¹¹ (**Supplementary Note 10**).

383 Altogether, granulocytes, smooth muscle and neurons showed more limited variation in
 384 mutation rate and spectra across individuals than has been observed in epithelia exposed to
 385 exogenous mutagens, such as skin³, colon⁶ (**Fig 2c**), bronchus³³ or bladder^{8,34}. This suggests
 386 that the rate of endogenous mutagenesis across individuals is modest, at least in the cohort
 387 studied. The observation of a linear accumulation of mutations in post-mitotic neurons, with
 388 similar burdens and signatures to some mitotically active tissues, suggest that dominant
 389 mutational processes observed across tissues may act independently of cell division.

390
 391



392

393 **Figure 3 | Mutation landscape in neurons and smooth muscle.** a-b, Substitution and indel accumulation
 394 per neuron throughout life; point estimate confidence intervals as in **Fig 2a**; grey area shows simple linear
 395 model 95% confidence intervals. Intercept and slope for substitutions: 210.5 [-26.9-448.0] and 17.1 [13.7-
 396 20.5] (20.0 [19.1-20.9] without intercept), respectively. Intercept and slope for indels: 45.9 [-10.2-102.0]
 397 and 2.5 [1.7-3.3] (3.1 [2.9-3.3] without intercept), respectively. c-d, Substitution and indel spectra in neurons
 398 from healthy and Alzheimer's disease donors; a description of each type of indel can be found in **Extended**

399 **Data Figure 5d. e**, Signature decomposition using granulocytes, colonic crypts, smooth muscle and neurons
400 substitution data. **f**, Substitution rates in the whole cohort for genes in quartiles of expression, showing
401 different types of substitutions and indels. Lines show Poisson 95% confidence intervals. **g**, Substitution
402 rates in transcribed and quiescent/heterochromatin DNA across different cell types. Lines show Poisson 95%
403 confidence intervals; the corresponding mutation spectra are shown in **Extended Data Fig 8a**). **h**, Indel rates
404 in the whole cohort for genes in quartiles of expression, showing different types of indels. Lines show
405 Poisson 95% confidence intervals. **i**, Contribution of signatures A, B and C in neurons. **j-k**, Substitutions
406 and indels per cell in smooth muscle from 10 donors spanning different ages; point estimate confidence
407 intervals and linear mixed model confidence intervals as in **Fig 2b**. Intercept and slope for substitutions:
408 239.3 [-211.5-653.9] and 20.7 [13.6-28.0] (24.5 [22.4-26.8] without intercept), respectively. Intercept and
409 slope for indels: 50.0 [2.6-97.2] and 1.3 [0.4-2.3] (2.2 [1.8-2.7] without intercept), respectively. **l-m**,
410 Substitution and indel spectra in smooth muscle. **n**, Exposure to signatures A, B and C in smooth muscle for
411 each donor and organ of origin. **o-p**, Substitution and indel accumulation per year across different cell types.

412

413

414 **Discussion**

415

416 Building on duplex sequencing and BotSeqS, we have developed a protocol with mutation-
417 detection error rates in single DNA molecules under 5 errors per billion sites. This error rate
418 enables the study of mutation rates and signatures in any human tissue or cell subpopulation.

419

420 Most of our current knowledge of somatic mutagenesis is restricted to mitotically-active cells.
421 We have exploited the ability to sequence any cell type to explore the mutational landscape of
422 non-dividing cells in a diversity of mitotically-active or inactive tissues. This has enabled us to
423 compare the mutational landscape of differentiated cells and stem cells in blood and colon, and
424 to study somatic mutagenesis in the absence of cell division. A remarkable observation that
425 emerges from these data is that somatic mutation rates vary modestly (~2-3 fold) across a
426 diverse range of somatic cell types, largely independently of cell division rates (**Fig 3o,p**,
427 **Suppl. Note 6**). Indeed, similar mutation rates are found in non-dividing cortical neurons, in
428 smooth muscle and in blood; or in colonic epithelium, which divides every few days, and in
429 mostly quiescent hepatocytes¹⁰ or urothelial cells (**Fig 3o,p**).

430

431 DNA replication and cell division have long been assumed to be major sources of somatic
432 mutations, either due to DNA polymerase errors or the fixation of unrepaired damage during
433 replication³⁵. However, the linear accumulation of somatic mutations in post-mitotic neurons
434 confirms that dominant mutational processes can occur independently of cell division. These
435 mutations may result from the interplay between endogenous DNA damage and repair that
436 cells are engaged in at all times. The similar mutation burden and signatures in granulocytes
437 and in the stem cells responsible for long-term maintenance of blood, despite a different
438 divisional load, could also be consistent with a time-dependent rather than a division-dependent
439 accumulation of somatic mutations during haematopoiesis. Altogether, it is conceivable that
440 division-independent mutational processes play a larger role in adult somatic mutagenesis than
441 it is commonly assumed.

442

443 In addition to enabling studies on somatic mutagenesis in any tissue, the ability to accurately
444 detect mutations in single molecules of DNA has wider applications. NanoSeq could be used
445 for mutagenesis screens and in vitro studies, exposing cell cultures or experimental models to
446 different mutagens and quantifying mutagenesis across the genome and over time, without the
447 need of single-cell bottlenecks^{36,37}. Sonication followed by exonuclease digestion opens the
448 door to targeted applications, to study the landscape of driver or pathogenic mutations from
449 polyclonal samples with reliable single-molecule detection, across tissues and conditions.
450 Being insensitive to clonality, NanoSeq can also be used to efficiently and accurately quantify

451 somatic mutation rates and signatures in liquid or non-invasive tissue samples, enabling studies
452 of somatic mutagenesis in large-scale cohorts, across genetic backgrounds, exposures and risk
453 factors, in health and disease.

454

455

456 **References**

457

458 1 Kennedy, S. R., Loeb, L. A. & Herr, A. J. Somatic mutations in aging, cancer and
459 neurodegeneration. *Mech Ageing Dev* **133**, 118-126, doi:10.1016/j.mad.2011.10.009
460 (2012).

461 2 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558,
462 doi:10.1126/science.1235122 (2013).

463 3 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection
464 of somatic mutations in normal human skin. *Science* **348**, 880-886,
465 doi:10.1126/science.aaa6806 (2015).

466 4 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age.
467 *Science* **362**, 911-917, doi:10.1126/science.aau3879 (2018).

468 5 Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal
469 expansion across normal tissues. *Science* **364**, doi:10.1126/science.aaw0726 (2019).

470 6 Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial
471 cells. *Nature* **574**, 532-537, doi:10.1038/s41586-019-1672-7 (2019).

472 7 Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic
473 human liver. *Nature* **574**, 538-542, doi:10.1038/s41586-019-1670-9 (2019).

474 8 Li, R. *et al.* Macroscopic somatic clonal expansion in morphologically normal human
475 urothelium. *Science* **370**, 82-89, doi:10.1126/science.aba7300 (2020).

476 9 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia.
477 *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

478 10 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells
479 during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

480 11 Franco, I. *et al.* Somatic mutagenesis in satellite cells associates with human skeletal
481 muscle aging. *Nat Commun* **9**, 800, doi:10.1038/s41467-018-03244-6 (2018).

482 12 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental
483 and transcriptional history. *Science* **350**, 94-98, doi:10.1126/science.aab1785 (2015).

484 13 Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased
485 mutations in single human neurons. *Science* **359**, 555-559,
486 doi:10.1126/science.aao4426 (2018).

487 14 Brazhnik, K. *et al.* Single-cell analysis reveals different age-related somatic mutation
488 profiles between stem and differentiated cells in human liver. *Sci Adv* **6**, eaax2659,
489 doi:10.1126/sciadv.aax2659 (2020).

490 15 Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines
491 Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294.e1220,
492 doi:10.1016/j.cell.2019.02.012 (2019).

493 16 Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation
494 sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**, 269-285,
495 doi:10.1038/nrg.2017.117 (2018).

496 17 Ahn, E. H. *et al.* Detection of Ultra-Rare Mitochondrial Mutations in Breast Stem
497 Cells by Duplex Sequencing. *PLoS One* **10**, e0136216,
498 doi:10.1371/journal.pone.0136216 (2015).

499 18 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing.
500 *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).

- 501 19 Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal
502 human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**,
503 9846-9851, doi:10.1073/pnas.1607794113 (2016).
- 504 20 You, X. *et al.* Detection of genome-wide low-frequency mutations with Paired-End
505 and Complementary Consensus Sequencing (PECC-Seq) revealed end-repair-derived
506 artifacts as residual errors. *Arch Toxicol* **94**, 3475-3485, doi:10.1007/s00204-020-
507 02832-0 (2020).
- 508 21 Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic
509 mutations. *Nature* **561**, 473-478, doi:10.1038/s41586-018-0497-0 (2018).
- 510 22 Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep
511 coverage targeted capture sequencing data due to oxidative DNA damage during
512 sample preparation. *Nucleic Acids Res* **41**, e67, doi:10.1093/nar/gks1443 (2013).
- 513 23 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to
514 disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).
- 515 24 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet*
516 **48**, 126-133, doi:10.1038/ng.3469 (2016).
- 517 25 Wyles, S. P., Brandt, E. B. & Nelson, T. J. Stem cells: the pursuit of genomic
518 stability. *Int J Mol Sci* **15**, 20948-20967, doi:10.3390/ijms151120948 (2014).
- 519 26 Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human
520 Colonic Epithelium. *Cell Stem Cell* **22**, 909-918.e908,
521 doi:10.1016/j.stem.2018.04.020 (2018).
- 522 27 Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by
523 genotoxic pks(+) E. coli. *Nature* **580**, 269-273, doi:10.1038/s41586-020-2080-8
524 (2020).
- 525 28 Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic
526 variation, and neurological disease. *Science* **341**, 1237758,
527 doi:10.1126/science.1237758 (2013).
- 528 29 Park, J. S. *et al.* Brain somatic mutations observed in Alzheimer's disease associated
529 with aging and dysregulation of tau phosphorylation. *Nat Commun* **10**, 3090,
530 doi:10.1038/s41467-019-11000-7 (2019).
- 531 30 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer.
532 *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).
- 533 31 Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole
534 genomes. *Nature* **578**, 102-111, doi:10.1038/s41586-020-1965-x (2020).
- 535 32 Gabella, G. Cells of visceral smooth muscles. *J Smooth Muscle Res* **48**, 65-95,
536 doi:10.1540/jsmr.48.65 (2012).
- 537 33 Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial
538 epithelium. *Nature* **578**, 266-272, doi:10.1038/s41586-020-1961-1 (2020).
- 539 34 Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in
540 the human bladder. *Science* **370**, 75-82, doi:10.1126/science.aba8347 (2020).
- 541 35 Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the Dependence of
542 Mutation Rates on Age and Time. *PLoS Biol* **14**, e1002355,
543 doi:10.1371/journal.pbio.1002355 (2016).
- 544 36 Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents.
545 *Cell* **177**, 821-836.e816, doi:10.1016/j.cell.2019.03.001 (2019).
- 546 37 Matsumura, S. *et al.* Genome-wide somatic mutation analysis via Hawk-Seq™
547 reveals mutation profiles associated with chemical mutagens. *Arch Toxicol* **93**, 2689-
548 2701, doi:10.1007/s00204-019-02541-3 (2019).

549 **Methods**

550

551 **Granulocytes and HSC/MPP colonies: sorting, colony growth and mutation calling**

552

553 We use two different terms to refer to colonies derived from haematopoietic stem cells (HSC)
554 or progenitor cells, depending on the membrane markers used for cell sorting: HSPCs, which
555 refer to CD34+ pools, and HSC/MPPs, which refer to CD34+ CD38- CD45RA- cells.

556

557 A sample of granulocytes from a 59-year-old male donor (PD43976_59yo) from whom 110
558 HSPC colonies were available²¹ was used for initial validation of the BotSeqS and NanoSeq
559 protocols (**Supplementary Tables 1,2**). To estimate the NanoSeq error rate, cord blood
560 granulocytes from two neonatal donors were sequenced by NanoSeq and the mutation burdens
561 and spectra compared to those from 50 HSC/MPP colonies per donor. For the comparison of
562 differentiated and stem cells, NanoSeq data from granulocytes from 9 donors of different ages
563 was compared to standard sequencing of single-cell derived HSC/MPP colonies from 6 donors
564 (10 HSC/MPP colonies per donor) and 110 HSPC colonies already available from a 59-year-
565 old donor²¹. These 110 HSPC included 67 HSC/MPPs, 32 megakaryocyte–erythrocyte
566 progenitors (MEP), 7 granulocyte–macrophage progenitors (GMP) and 4 common myeloid
567 progenitors (CMP).

568

569 For PD43976_59yo, HSPC colonies were grown and mutations called as described in Lee Six
570 *et al.*²¹. For the remaining donors, whole blood was diluted with PBS and mononuclear cells
571 (MNC) were isolated using lymphoprepTM (STEMCELL Technologies) density gradient
572 centrifugation. The red blood cell and granulocyte fraction of the blood was then removed. The
573 MNC fraction was depleted of red blood cells by lysis steps involving 3 incubations at room
574 temperature for 20 mins/10 mins/10 mins respectively with RBC lysis buffer (BioLegend).
575 CD34+ selection of peripheral blood and cord blood samples was undertaken using the
576 EasySep human whole blood CD34 positive selection kit (STEMCELL Technologies) as per
577 the manufacturer’s instructions. Bone marrow samples did not undergo CD34+ selection prior
578 to sorting.

579

580 MNC or CD34 enriched samples were centrifuged and resuspended in PBS/3%FBS containing
581 an antibody panel consisting of (antibody/fluorochrome): CD3/FITC, CD90/PE,
582 CD49f/PECy5, CD38/PECy7, CD19/A700, CD34/APC Cy7, CD45RA/BV421, and
583 Zombie/Aqua.

584

585 Cells were stained (30 minutes at 4°C) in the dark before washing, centrifugation (500 x g at
586 room temperature) and resuspension in PBS/3%FBS for cell sorting. Index sorting of
587 ‘HSC/MPP pool’ cells was performed on a BD AriaIII Cell Sorter (BD Biosciences) at the
588 NIHR Cambridge BRC Cell Phenotyping Hub, as per the gating structure in **Extended Data**
589 **Fig 6a** (CD34+, CD38- and CD45RA-).

590

591 ‘HSC/MPP pool’ cells were single-cell sorted into Nunc 96 well flat-bottomed TC plates
592 (ThermoFisher) containing 100 µl supplemented StemPro media (Stem Cell Technologies).
593 MEM media contained StemPro Nutrients (0.035%, Stem Cell Technologies), L-Glutamine
594 (1%, ThermoFisher), Penicillin-Streptomycin (1%, ThermoFisher) and cytokines (SCF, 100
595 ng/ml; FLT3, 20 ng/ml; TPO, 100 ng/ml; EPO 3 ng/ml; IL-6, 50 ng/ml; IL-3, 10 ng/ml; IL-11,
596 50 ng/ml; GM-CSF, 20 ng/ml; IL-2 10 ng/ml; IL-7 20 ng/ml; lipids 50 ng/ml) to promote
597 differentiation towards Myeloid/Erythroid/Megakaryocyte (MEM) and NK lineages. Manual
598 assessment of colony growth was made at 14 days. Colonies were topped up with an additional

599 50 µL MEM media on day 15 if the colony was $\geq 1/4$ size of well. Following 21 ± 2 days in
600 culture, colonies were selected by size criteria. Colonies ≥ 3000 cells in size were harvested
601 into a U bottomed 96 well plate (ThermoFisher). Plates were then centrifuged (500 x g for 5
602 minutes), media was discarded, and the cells were resuspended in 50 µl PBS prior to freezing
603 at -80°C . Colonies < 3000 cells but > 200 cells in size were harvested into 96 well skirted
604 LoBind plates (Eppendorf) and centrifuged (800 x g for 5 min). Supernatant was removed to
605 5-10 µL using an aspirator prior to DNA extraction on the fresh cell pellet.

606
607 DNA extraction was performed using the DNeasy 96 blood and tissue plate kit (Qiagen) for
608 larger HSC colonies, or the Arcturus Picopure DNA Extraction kit (ThermoFisher) for smaller
609 HSC colonies. Both kits were used as per the manufacturer's instructions. Extracted DNA (1-
610 5ng) from each colony was processed using a recently developed low-input enzymatic
611 fragmentation-based library preparation method³⁸. All samples were subjected to whole
612 genome sequencing at 8-35X coverage on either the HiSeq X or the NovaSeq platforms
613 (Illumina) to generate 150 bp paired-end reads. BWA *mem* was used to align sequences to the
614 human reference genome (NCBI build37).

615
616 The haematological samples in the study were obtained from several sources: the Cambridge
617 Blood and Stem Cell Biobank, the Cambridge Biorepository for Translational Medicine, and
618 the Cambridge Bioresource (REC references: 07-MRE05-44, 18/EE/0199, 15/EE/0152 -
619 NRES Committee East of England - Cambridge South).

620

621 **Sperm samples**

622

623 DNA was extracted from sperm samples from two donors, aged 21 and 73 years, and sequenced
624 using the NanoSeq protocol (REC ethics approval: EC04/015, London - Westminster REC;
625 16/NE/003, NRES Committee North East-Newcastle and North Tyneside 1). Because of the
626 low mutation burden of the germline, we sequenced 7 separate aliquots of sperm DNA from
627 the 21-year-old donor to estimate the error rate of the NanoSeq protocol (**Supplementary**
628 **Tables 1,2**).

629

630 **Laser microdissection of colonic crypts and bladder/colon smooth muscle**

631

632 Colon and bladder biopsies were obtained from deceased organ donors (ranging in age from
633 25 to 78; **Supplementary Table 1**) at the time of organ donation. This tissue was collected as
634 part of the Cambridge Biorepository for Translational Medicine program (REC reference:
635 15/EE/0152 NRES Committee East of England – Cambridge South). Families of the donors
636 provided informed consent for the use of this material in research. Different microbiopsies
637 from these specimens have been used in previously published studies^{6,34,39}.

638

639 Colon biopsies were fresh frozen at the time of collection and stored at -80°C . The colon
640 biopsies subsequently underwent formalin-free fixation for 24 hours in PAXgene Tissue Fix
641 containers (PreAnalytiX, Hombrechtikon, Switzerland) before being transferred to PAXgene
642 STABILIZER solution (PreAnalytiX). Bladder biopsies underwent formalin-free fixation at
643 the time of collection and were stored at -20°C ³⁸.

644

645 Prior to laser-capture microdissection, samples were processed, embedded in paraffin and
646 sectioned as described previously³⁴. Microbiopsies were dissected using an LMD7 microscope
647 (Leica Microsystems). Examples of microdissected regions for both specimen types can be
648 found in **Extended Data Figures 6 and 10**. Proteolysis of isolated regions was performed using

649 an Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific, Waltham, MA, USA).
650 Cell lysate was stored at -20 °C prior to library preparation.

651

652 **Neuron nuclei sorting from frontal cortex samples**

653

654 Frozen biopsies of frontal cortex from eight healthy and nine Alzheimer's disease donors were
655 collected by the Cambridge Brain Bank (**Supplementary Tables 1,2**; REC ethics approval:
656 10/H0308/56, East of England, Nottingham). Neuronal nuclei were isolated, stained and
657 extracted from the frontal cortex samples as per Krishnaswami et al.⁴⁰. Briefly, small cuts of
658 1-2 mm were taken from fresh frozen samples. Dounce homogenisation was then used to free
659 nuclei before filtration, density centrifugation and immunostaining. Samples were stained
660 using DAPI (Thermo Fisher, D1306) and Milli-Mark™ Anti-NeuN-PE Antibody (MilliPore,
661 FCMAB317PE). The immunostained samples were then sorted using FACS as per the gating
662 strategy in **Extended data Fig 7a**. 15,000 nuclei were collected into 20 µl Arcturus PicoPure
663 DNA Extraction Kit (Thermo Fisher Scientific) before undergoing digestion. Nuclear lysate
664 was then stored at -20°C prior to library preparation.

665

666 The distributions of NeuN-PE intensities in most samples revealed a bimodal distribution. As
667 a quality control, we fitted a mixture of two Gamma distributions to the NeuN-PE intensities
668 for every samples. Only samples with 10-fold (1 log₁₀ unit) separation between the mean of
669 both peaks were considered for analysis, which led to the exclusion of an outlier sample.

670

671 **BotSeqS and NanoSeq library preparation protocols**

672

673 BotSeqS libraries were prepared as follows: DNA was fragmented using focused
674 ultrasonication (Covaris 644 LE220) and purified by 2.5x AMPure XP (Beckman Coulter). 10
675 ng of sonicated DNA was end-repaired and ligated using the NEBNext Ultra II kit (New
676 England Biolabs) including 0.66 µl 1.5 µM xGen Duplex Seq Adapters - Tech Access
677 (Integrated DNA Technologies, IDT: 1080799).

678

679 NanoSeq libraries were prepared as follows: 10 ng of genomic DNA or LCM cut sections in
680 20 µl buffer were purified using 100 µl of a 50:50 water and AMPure XP bead mixture and
681 eluted in 20 µl nuclease free water. 20 µl of the bead suspension was taken forward into an on-
682 bead fragmentation reaction. Fragmentation occurred in a final volume of 25 µl including 2.5
683 µl 10x CutSmart buffer (500 mM Potassium Acetate, 200 mM Tris-acetate, 100 mM
684 Magnesium Acetate, 1 mg/ml BSA, pH 7.9 at 25°C), 0.5 µl 5 U/µl HpyCH4V (**Supplementary**
685 **Note 2**), and 2 µl dH₂O. Fragmentation reactions were incubated at 37 °C for 15 min, purified
686 with 2.5x AMPure XP beads and resuspended in 15 µl nuclease-free water. Fragmented DNA
687 was A-tailed in 15 µl reactions including 10 µl fragmentation product, 1.5 µl 10x NEBuffer 4
688 (500 mM Potassium Acetate, 200 mM Tris-acetate, 100 mM Magnesium Acetate, 10 mM DTT,
689 pH 7.9 at 25°C), 0.15 µl 5 U/µl Klenow fragment (3'→5' exo-, New England Biolabs), either
690 1.5 µl 1 mM dATP or 1.5 µl 1 mM dATP/ddBTPs (**Supplementary Note 3**), and 1.85 µl
691 dH₂O. Reactions were incubated at 37 °C for 30 mins. The 15 µl A-tailing reaction product was
692 added to 22.4 µl ligation mix, which consisted of 2.24 µl 10x NEBuffer 4, 3.74 µl 10 mM ATP,
693 0.33 µl 15 µM xGen Duplex Seq Adapters (IDT: 1080799), 0.56 µl 400 U/µl T4 DNA ligase
694 (New England Biolabs), and 15.53 µl dH₂O. Reactions were incubated at 20 °C for 20 min and
695 subsequently purified with 1x AMPure XP beads and resuspended in 50 µl of nuclease free
696 water.

697

698 DNA quantification, dilution and PCR amplification

699

700 DNA was quantified by qPCR using a KAPA library quantification kit (KK4835). The supplied
701 primer premix was first added to the supplied KAPA SYBR FAST master mix. In addition, 20
702 μl of 100 μM NanoqPCR1 primer (HPLC: 5'-ACACTCTTCCCTACACGAC-3') and 20 μl
703 of 100 μM NanoqPCR2 primer (HPLC: 5'-GTGACTGGAGTTCAGACGTG-3') were added
704 to the KAPA SYBR FAST master mix. Samples were diluted 1 in 500 using nuclease-free
705 water and reactions were set up in a 10 μl reaction volume (6 μl master mix, 2 μl
706 sample/standard, 2 μl water) in a 384 well plate. Samples were run on the Roche 480
707 Lightcycler and analysed using absolute quantification (2nd Derivative Maximum Method)
708 with the high sensitivity algorithm. nM (fmol/ μl) was determined as follows: mean of sample
709 concentration x dilution factor (500) x 452/573/1000 (where 452 is the size of the standard in
710 bp and 573 is the proxy for the average fragment length of the library in bp), and multiplied by
711 an adjustment factor of 1.5. Samples were diluted to the desired fmol amount (typically 0.3
712 fmol for a 15x run) in 25 μl using nuclease free water.

713

714 Libraries were subsequently PCR amplified in a 50 μl reaction volume comprising of 25 μl
715 sample, 25 μl NEBNext Ultra II Q5 Master Mix and UDI containing PCR primers (dried). The
716 reaction was cycled as follows: step1: 98 $^{\circ}\text{C}$ 30 seconds, step2: 98 $^{\circ}\text{C}$ 10 seconds, step3: 65 $^{\circ}\text{C}$
717 75 seconds, step4: return to step2 13 times, step5: 65 $^{\circ}\text{C}$ for 5 min, step6: hold at 4 $^{\circ}\text{C}$. The
718 number of PCR cycles is dependent upon the input: 0.1 fmol = 16 cycles, 0.3 fmol = 14 cycles,
719 0.6 fmol = 13 cycles, 5 fmol = 10 cycles.

720

721 The PCR product was subsequently cleaned up using two consecutive 0.7x AMPure XP clean-
722 ups. Each sample was quantified using the AccuClear Ultra High Sensitivity dsDNA
723 Quantification kit (Biotium) and pooled. Libraries were sequenced on Illumina sequencing
724 platforms e.g. NovaSeq using 150 paired-end reads.

725

726 Library dilution and sequencing efficiency

727

728 The efficiency and cost-effectiveness of duplex sequencing depends on optimising the
729 duplicate rate to maximise the number of read bundles (defined as a family of PCR duplicates)
730 with at least 2 duplicate reads from each original strand. Too high duplicate rates result in few
731 read bundles of unnecessarily large sizes, whereas too low duplicate rates result in many read
732 bundles with few having two or more read pairs from each strand.

733

734 To maximise the efficiency of the protocol, we studied analytically and empirically the
735 relationship between the number of DNA molecules in the library (library complexity) and the
736 resulting duplicate rate as a function of the number of read pairs sequenced. We found that
737 optimal duplicate rates and optimal efficiency can be ensured across a wide range of samples.
738 If we assume negligible PCR biases, with copies from all original ligated DNA fragments
739 represented in equimolar amounts in the amplified library, the bundle size distribution of
740 observed reads can be modelled as a zero-truncated Poisson distribution. Let r (sequence ratio)
741 be the ratio between the number of sequenced reads and the number of amplifiable DNA
742 fragments in the original library. The mean read bundle size (m) can then be estimated as the
743 mean of the zero-truncated Poisson distribution: $m = \frac{r}{1-e^{-r}}$. This parameter then enables a
744 simple estimation of the duplicate rate of a library (d , defined as the fraction of reads that are
745 duplicate copies, and identified as reads having the same barcodes and the same 5' and 3'
746 coordinates): $d = \frac{m-1}{m} = 1 - \frac{1}{m} = 1 - \frac{1-e^{-r}}{r}$.

747
748 We can define the efficiency of a duplex sequencing library (E) as the ratio between the number
749 of base pairs with duplex coverage (bundles with ≥ 2 reads from both strands) and the number
750 of base pairs sequenced. This can be modelled as: $E = \frac{P(x \geq 2; \frac{r}{2})^2}{m}$, where the numerator is the
751 probability of a read bundle having at least two reads from both strands (i.e. usable bundles),
752 based on the zero-truncated Poisson distribution (denoted as P), and the denominator is the
753 sequence investment in each read bundle (i.e. the average read bundle size). Based on this
754 equation, we can estimate numerically that the optimal duplicate rate is $\sim 81\%$ (**Extended Data**
755 **Fig 4a, Supplementary Code**) and that duplicate rates between 65-90% would yield $\geq 80\%$ of
756 the maximum attainable efficiency. In terms of r , the optimum r is 5.1 read pairs sequenced
757 per original DNA fragment (r_{opt}), with values within 2.7-9.6 yielding $\geq 80\%$ of the maximum
758 efficiency. Knowing the concentration of a NanoSeq (or BotSeqS) library in fmol/ μ l (estimated
759 using a qPCR reaction on an aliquot of the library), we can use r_{opt} to calculate the volume of
760 library that needs to be amplified to yield optimal duplicate rates (i.e. maximum duplex
761 efficiency), as a function of the desired amount of raw sequencing: $fmol_{opt} = \frac{N}{f r_{opt}}$. Here, N
762 is the number of paired-end reads that will be sequenced and f is the number of DNA fragments
763 per fmol of library (referring specifically to ligated and amplifiable fragments within the size
764 selection range). Using an initial set of libraries, we compared a range of library inputs (fmol)
765 to the estimated number of unique molecules in the library inferred from the sequencing data
766 (using Piccard's software). This analysis revealed that, for our choice of restriction enzyme
767 and size selection conditions, f approximately equated to 10^8 fragments/fmol (**Supplementary**
768 **Code**).

769
770 Using the above equation, we can optimise the efficiency of NanoSeq independently of the
771 input amount of DNA in a given sample. For example, ~ 0.3 fmols of library yield optimal
772 duplicate rates when using 150 million 150 bp paired-end reads, which are the equivalent of
773 $\sim 15x$ coverage in standard human whole-genome sequencing. ~ 0.6 fmol yield optimal
774 efficiency when using 300 million reads ($30x$ whole-genome equivalent). Note that, as
775 predicted by the equations above, deviations ~ 2 -fold from r_{opt} still yield high efficiency. Using
776 these equations we reliably obtained near-optimal duplicate rates from a wide diversity of
777 samples (**Extended Data Fig 4, Supplementary Table 2**). Overall, we found that $\sim 30x$ of
778 standard sequencing output ($\sim 300 \times 10^6$ 150bp PE reads) yielded approximately 3 Gb of high-
779 accuracy duplex coverage (a haploid genome equivalent) after application of all computational
780 filters.

781
782 Our choices of restriction enzyme and size selection restrict the coverage to $\sim 30\%$ of the human
783 genome. Although the covered regions are sufficiently diverse to enable unbiased estimates of
784 burden and signatures (**Methods**), applications that require full genome coverage, such as
785 targeted sequencing, would require alternative fragmentation strategies. One option may be
786 exonuclease blunting after sonication, instead of end repair. Nevertheless, for the study of
787 burden and signatures, the use of restriction enzymes has two interesting advantages. First, this
788 protocol is able to work with very low inputs of DNA. We estimated library yields for a range
789 of input DNA amounts (**Extended Data Fig 4b**) and found that the minimum DNA input
790 required to obtain 0.3 fmol for a $15x$ run (corresponding to about 1.5-3 Gb of effective duplex
791 coverage) was ~ 1 ng of input DNA. This low-input requirement enables the application of
792 NanoSeq to microscopic areas of tissue (as shown for colonic crypts and smooth muscle) and
793 to rare cell populations using flow sorting. A second advantage is that, since coverage is
794 concentrated in $\sim 30\%$ of the human genome, matched normal samples can be sequenced at

795 lower cost by using undiluted NanoSeq libraries (≥ 3 fmol of library sequenced at 8x genome
796 equivalent is enough to provide high matched normal coverage in the 30% of informative
797 genome).
798

799 **Sequencing, preprocessing and filtering of BotSeqS and NanoSeq libraries**

800
801 Standard sequencing matched-normal libraries were aligned to the human reference genome
802 (GRCh37, hs37d5 build) using BWA-MEM v0.7.5a-r405⁴¹ with default parameters.
803 Alignments were sorted by coordinate and read duplicates were marked using biobambam2⁴²
804 v2.076 bamsormadup. Matched-normal reads were filtered if marked as duplicate,
805 supplementary, QC fail, unmapped or secondary alignments. For some samples, as described
806 above, instead of standard whole-genome sequencing, we used undiluted NanoSeq libraries
807 (typically ~ 5 fmol) as matched normals, reducing the costs of sequencing matched normal
808 samples.
809

810 NanoSeq and BotSeqS libraries were sequenced using 150 bp paired-end reads, in HiSeq2500,
811 HiSeqX and NovaSeq platforms.
812

813 NanoSeq sequencing reads begin with adapter sequences: NNNT or NNNXT for BotSeqS
814 libraries and NNNTCA or NNNXTCA for HpyCH4V libraries (HpyCH4V cuts at TGCA
815 motifs). NNN is a random three nucleotide barcode, T is the adapter overhang and X is a
816 ‘spacer’ nucleotide designed to increase nucleotide diversity in the sequencing run. We used a
817 custom Python script to process demultiplexed fastq files by extracting the three-nucleotide
818 barcode, clipping remaining adapter bases (2 bases for BotSeqS and 4 bases for NanoSeq
819 libraries) and appending barcode sequences to the fastq header. Barcodes with non-canonical
820 bases (not A, C, G or T) were filtered out. Reads were aligned to hs37d5 using *bwa mem*
821 (v0.7.5a-r405), using the *-C* option to append barcode sequences to alignments. Alignments
822 were sorted by coordinate, duplicates were marked, and reads were annotated with read
823 coordinate, mate coordinate and optical duplicate auxiliary tags using biobambam2 v2.076
824 bamsormadup and bammarkduplicatesopt (optminpixeldif=2500). Reads were filtered when
825 they were not marked as proper-pairs or were marked as optical duplicate, supplementary, QC
826 fail, unmapped or secondary alignments. Each read was marked with an auxiliary tag
827 comprised of reference name, sorted read and mate fragmentation breakpoints, forward and
828 reverse read barcodes, and read strand.
829

830 **Consensus base quality scores**

831
832 Bayes’ theorem was used to compute the posterior probability of each base call B given the
833 pileup of reads D from one strand of a template molecule at one genomic position. There are
834 four possible genotypes $i \in (A, C, G, T)$. The posterior probability is calculated using:
835

$$836 \quad P(B|D) = \frac{P(B)P(D|B)}{\sum_i P(B_i)P(D|B_i)}$$

837
838 Under a uniform prior, where any of the four possible genotypes are equally likely, the equation
839 can be simplified to:
840

$$841 \quad P(B|D) = \frac{P(D|B)}{\sum_i P(D|B_i)}$$

842
843
844
845

To calculate $P(D|B)$, information is integrated from reads in D , where $b_j \in (A, C, G, T)$ is the base of read $j = 1 \dots d$:

$$P(D|B_i) = \prod_{j=1}^{j=d} P(b_j|B_i)$$

847
848
849
850

To calculate $P(b_j|B_i)$ we use the probability that base b_j is an error, calculated from its Phred quality score q_j :

851
852

$$P(b_j|G_i) = 1 - e_j \text{ if } b_j = B_i, \text{ otherwise } e_j/3$$

853
854

where

855
856

$$e_j = 10^{-\frac{q_j}{10}}$$

857
858
859
860

We note that the final probability $P(D|B)$ is the probability that the base call is correct after sequencing and not the probability that the base represents the correct genotype of the original template strand, where independence between observations cannot be assumed. $P(B|D)$ is rescaled into a Phred quality score Q using:

861
862
863

$$Q = -10 \log_{10} P(B|D)$$

864
865
866

In cases where the two read mates overlap, the consensus base quality is calculated using both forward and reverse reads.

867 **Base calling and filtering**

868
869
870
871
872

We developed a set of filters that successfully reduced false positive calls. An important feature of the bioinformatic pipeline is that we apply the same filters to call reference and mutated bases, which allows direct calculation of mutation rates.

873
874
875
876

The calling method requires a matched normal to filter out germline SNPs. An additional mask to filter sites that are problematic is also advisable. This matched normal can be obtained by standard protocols or by sequencing undiluted NanoSeq libraries (≥ 3 fmol), as explained above.

877
878
879

The filters applied are the following:

880
881
882
883
884
885
886

1. We require that each read bundle (i.e. group of PCR duplicates) has at least two reads from each of the two original DNA strands.
2. The consensus base quality score should be at least 60. This guarantees that there is strong support for a given base call from the duplicate reads that form a read bundle.
3. The minimum difference between the primary alignment score (AS) and the secondary alignment score (XS) should be higher than 50, to keep only read pairs with unambiguous mapping. This filter is essential to remove mapping artefacts and a

887 minimum AS-XS of 50 is applied also to the matched normal. For sites where the two
888 mates overlap the minimum of the average AS-XS for forward and reverse mates is
889 taken.

- 890 4. The average number of mismatches in a group of reads (forward or reverse) should not
891 be higher than 2. This filter is important to exclude reads with unreliable mappings.
892 Where a consensus base call is different from the reference, mismatches from this call
893 are not consider when calculating the number of mismatches, hence avoiding a bias in
894 the filtering of mutation and reference calls. This filter is also applied to the matched
895 normal. For sites where the two mates overlap the maximum of the average NM for
896 forward and reverse mates was taken.
- 897 5. No 5' clips are allowed.
- 898 6. No improper pairs are allowed in the read bundle to avoid unreliable mappings.
- 899 7. Base calls in read ends, defined as those within 8 bp from the 5' or 3' ends, are discarded
900 because these regions are more likely to be unreliably mapped, especially when there
901 are nearby indels.
- 902 8. Reads in the read bundle must contain no indels (except for indel calling).
- 903 9. The matched normal must have $\geq 15x$ coverage at a given site to make the risk of
904 undetected heterozygous SNPs negligible. For non-neat matched normals we also
905 require that there are at least 5 reads from each strand.
- 906 10. When a mutation is to be called, we require that the base is not seen with a frequency
907 higher than 0.01 in the matched normal. This filter is not applied when counting
908 reference calls, but we have assessed that our results are stable with different thresholds.
- 909 11. Finally, a site should not overlap the common SNP and noisy sites masks (see **Genome**
910 **masks**). Base calls failing this requirement are also counted to obtain a qualitative
911 diagnostic of potential contamination of the input DNA with DNA from a different
912 individual. In the presence of contamination, mutation rates can be considerably
913 inflated if these masks are not applied.

914

915 **Indel calling**

916

917 To call indels we first identify read bundles with potential indels, defined as those containing
918 sites with at least 90% of forward and reverse reads having an indel. Read bundles with AS-
919 XS ≤ 50 , 5' clipping or with coverage in the matched normal lower than 16 were filtered out.
920 Indels close to read ends (10 bp) were not called. For each of the read bundles potentially
921 containing an indel, the corresponding reads were extracted from the BAM file, removing PCR
922 duplicate flags and creating a mini read bundle BAM. For each of the read bundle BAMs we
923 run samtools mpileup to generate genotype likelihoods in BCF format, as follows:

924

```
925 samtools mpileup --no-BAQ -d 250 -m 2 -F 0.5 -r $chr:$start-$end --BCF --output-  
926 tags DP,DV,DP4,SP -f $ref_genome -o genotype_likelihods.bcf read_bundle.bam
```

927

928 where \$chr, \$start and \$end are the mapping coordinates of the read bundle. Next, we call
929 indels and normalise the output using bcftools as follows:

930

```
931 bcftools index -f genotype_likelihods.bcf genotype_likelihods.indexed.bcf
```

932

```
933 bcftools call --skip-variants snps --multiallelic-caller --variants-only -O v  
934 genotype_likelihods.bcf -o bcftools.tmp.vcf
```

935

```
936 bcftools norm -f $ref_genome bcftools.tmp.vcf > bcftools.tmp2.vcf
```

937
938 For each of the sites involved in an indel we check whether it overlaps a site masked by our
939 common SNP and noise masks (see **Genome masks**), in which case the indel is flagged as
940 MASKED and not further analysed.

941
942 The final step involves revisiting the matched normal to inspect if there are indels in a window
943 of ± 5 bp around each candidate indel. For this step we use the bam2R function from R package
944 *deepSNV*⁴³. Reads with mapping quality lower than 10 or with any of the following flags are
945 ignored: "read unmapped", "not primary alignment", "read fails platform/vendor quality
946 checks", "read is PCR or optical duplicate", and "supplementary alignment". If the proportion
947 of indels in the matched normal within the ± 5 bp window around the candidate somatic indel
948 is higher than 1%, the indel is disregarded.

949

950 **Substitution imbalances**

951

952 To detect asymmetries in substitution patterns, variants were assigned to the forward or reverse
953 strand according to their distance from fragmentation breakpoints. Variants closest to the 5' of
954 the forward read were assigned to the forward strand. Variants closest to the 5' of the reverse
955 read were assigned to the reverse strand and reverse complemented. Variants equidistant from
956 both fragmentation breakpoints were not counted.

957

958 **Genome masks**

959

960 We applied two masks to filter duplex sequencing data. The first mask comprised common
961 SNPs and spanned a total of 27,204,965 bp. Autosomal and X-chromosome common SNPs
962 were defined as SNPs with allele frequency (AF) $> 0.1\%$ and a "PASS" flag in gnomAD. Y-
963 chromosome and mitochondrial SNPs were defined as SNPs with AF $> 0.1\%$ from 1000
964 Genomes Project (1KGP) data^{44,45}. This SNP mask is important to reduce the impact of
965 potential inter-individual DNA contamination (**Supplementary Note 6**).

966

967 A second mask was developed to remove unreliable calls or sites prone to alignment artefacts.
968 To build this noise mask we gathered together gnomAD indel calls with AF $> 1\%$ and SNP calls
969 with AF $> 0.1\%$ that were not flagged as "PASS". The noise mask also contains sites with
970 elevated error-rates. To generate it, mismatch rates were calculated for every genomic position
971 across a panel of 448 in-house standard whole-genome samples. Sites with mismatch rates
972 (coverage-weighted mean VAF) > 0.01 were incorporated into the noise mask. Altogether, the
973 second mask comprised 22,474,160 bp.

974

975 Both masks are available at https://github.com/fa8sanger/NanoSeq_Paper_Code.

976

977 **Detection of human DNA contamination**

978

979 Contamination of duplex sequencing libraries with DNA from other individuals could
980 artificially inflate mutation burden estimates, mainly because germline SNPs in the
981 contaminant DNA may appear as somatic mutations.

982

983 Even a small percentage of contamination can have a large impact on burden estimates. The
984 burden associated to SNPs in the contaminant would be:

985

986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033

$$Burden_{SNP} = \frac{N_{SNP} * f_{cont}}{G}$$

being N_{SNP} the number of SNPs in the contaminant not shared with the sample at hand, f_{cont} the contamination fraction and G the size of the diploid human genome. Accordingly, 1% contamination would result in a $Burden_{SNP}$ of $\sim 5 \times 10^{-6}$ if there are 3 million non-shared SNPs. This burden is much higher than the usually observed somatic mutation rates.

First, we analysed how many SNPs across 2,504 individuals from the 1000 Genomes Project would remain after filtering with our common SNPs mask ($n=26,111,286$; **Methods**). Our results show that on average 55,685 SNPs would remain unfiltered for a given contaminant individual. Hence, for 1% contamination, filtering of common SNPs would reduce $Burden_{SNP}$ from 5×10^{-6} to 9×10^{-8} SNPs/bp. We note that the number of unfiltered SNPs varies largely across continental groups, with averages of 25,666 and 82,765 per individual in Europe and South Asia, respectively (**Supplementary Note 6**).

To estimate the extent of contamination we rely on VerifyBamID2⁴⁶, which we evaluated simulating contamination fractions below 1%, for both bam files sequenced with standard methods and with the NanoSeq protocol (**Extended Data Fig 4e,f**; **Supplementary Note 6**). To obtain more stable estimates we increased the number of markers from 100K to 500K, by randomly choosing additional SNPs with $MAF > 0.05$ from the 1000 Genomes Project 20130502 release.

***In silico* decontamination**

We detected that some libraries were contaminated with DNA from other analysed samples. In cases where the contaminant can be identified, it is possible to remove the mutation calls corresponding to contaminant SNPs by using the corresponding BAM files. This simple approach proved useful to clean contaminated substitution calls and resulting mutation burden corrections were in line with VerifyBamID contamination estimates. That is, mutation burdens of non-contaminated samples remained unaltered after *in silico* decontamination, whereas the mutation burdens of contaminated samples decreased proportionally to the estimated contamination level.

This approach was applied to two plates where some samples showed signs of contamination. Mutation calls occurring at SNP sites in any of the other samples in the plate were removed. To accomplish this we required that each mutation was supported by fewer than 10 base calls across the matched normals of samples in the plate and that the maximum support from any one matched normal sample was lower than 3 reads. These values were found empirically for the data at hand and should be adjusted when larger panels of matched normals or very high coverage samples are analysed.

Correction of mutation burden and trinucleotide substitution profiles

Each library preparation method has its own fragmentation and amplification biases and captures a different subset of the total genome. For instance, amplification biases during library preparation often lead to lower coverage in GC-rich genomic regions⁴⁷. Since substitution rates show strong trinucleotide context dependence, taking into consideration differences in sequence composition can be important when comparing mutation burdens and substitution profiles between sequencing protocols. Biases can be particularly noticeable with NanoSeq

1034 restriction enzyme libraries, where trinucleotides overlapping the restriction enzyme site
 1035 (TGCA in the case of HpyCH4V) are depleted when read ends are filtered. There are 32
 1036 different trinucleotides where the central nucleotide is a pyrimidine. Let t denote the count of
 1037 a given trinucleotide of type $i = 1...32$. The frequency of each trinucleotide is calculated
 1038 separately for the genome f_i^g and for the NanoSeq experiment (weighted by the coverage at
 1039 each site) f_i^e where:

$$f_i = \frac{t_i}{\sum_{i=1}^{32} t_i}$$

1041

1042

1043 The ratio of genomic to experimental frequencies for a given trinucleotide is:

1044

$$r_i = \frac{f_i^g}{f_i^e}$$

1045

1046

1047 There are six classes of substitution where the mutated base is a pyrimidine (C>A, C>G, C>T,
 1048 T>A, T>C, T>G), and for each trinucleotide context there are three possible substitutions. Each
 1049 trinucleotide-substitution count (e.g. ATG>C, where T>C) is corrected by the ratio of genomic
 1050 to experimental frequencies for the corresponding trinucleotide (ATG). For instance, let $s_{ATG>C}$
 1051 denote the count of substitution $T>C$ in trinucleotide context ATG , the substitution count is
 1052 corrected as follows:

1053

1054

$$s'_{ATG>C} = s_{ATG>C} r_{ATG}$$

1055

1056 This correction is applied to each of the 96 possible trinucleotide substitutions (h). The
 1057 corrected substitution counts provide a substitution profile projected onto the human genome,
 1058 and are also used to calculate the corrected mutation burden:

1059

$$\beta' = \frac{\sum_{h=1}^{96} s'_h}{\sum_{i=1}^{32} t_i}$$

1060

1061

1062

1063 **Correction of NanoSeq mutation burden in cord blood by accounting for missed early** 1064 **embryonic mutations**

1065

1066 Given their low burden, a substantial fraction of the mutation burden in cord blood HSC/MPP
 1067 colonies is attributable to early embryonic mutations shared by multiple colonies. In the
 1068 NanoSeq bioinformatic protocol, mutations with a VAF higher than 0.01 in the matched normal
 1069 are considered germline SNPs and are filtered out from further analysis. Not accounting for the
 1070 loss of early embryonic mutations can have a measurable impact on burden estimates in cord
 1071 blood. Taking advantage of the availability of multiple HSC/MPP colonies per donor, we could
 1072 quantify the loss of embryonic variants and correct the burden estimate accordingly. For each
 1073 of the 50 blood colonies we estimated the global VAF of each mutation in the remaining 49
 1074 colonies. This was done for the two neonatal donors. We determined that 24% of all the
 1075 mutations called had a global VAF higher than 0.01. Since a similar fraction of mutations
 1076 would be missed by NanoSeq, we multiplied the NanoSeq estimated burden by a factor of 1.32,
 1077 i.e. $1/(1-0.24)$. A similar correction is not possible for the sperm burden estimates, as we lack

1078 single-cell level information for sperm, but a modest underestimation of the mutation burden
1079 due to missed embryonic variants is plausible.

1080

1081 **Mutation calling in clonal samples sequenced with standard protocols**

1082

1083 Mutation calls for HSPC colonies from donor PD43976_59yo were obtained from Lee-Six *et al.*
1084 *et al.* 2018²¹. Mutation calls from standard whole-genome sequencing for the colonic crypts
1085 processed in Lee-Six *et al.* 2019⁶ were obtained from Olafsson *et al.*³⁹. Indel mutation calls for
1086 a bladder tumour sample (**Extended Data Fig 5**) were obtained from Lawson *et al.*³⁴. Indel
1087 calls for POLE and POLD1 mutants were obtained from Robinson *et al.*⁴⁸ (**Extended Data Fig**
1088 **5**).

1089

1090 For the HSC/MPP blood colonies sequenced in the present study, in-house pipelines were used
1091 to run CaVEMan and Pindel against an unmatched synthetic normal genome^{49,50}. Another
1092 bespoke algorithm (cgpVAF) was then used to generate matrices of variant and normal reads
1093 at all sites that had a detected variant in any sample from a given individual. Up-to-date
1094 versions of these algorithms are available from the Sanger Institute's Cancer IT GitHub
1095 repository (<https://github.com/cancerit>).

1096

1097 Filtering strategies detailed below were then used to remove germline variants, technical
1098 artefacts and mutations that had arisen during culture in vitro.

1099

- 1100 1. A custom filter was used to remove artefacts associated with the 'low input' library
1101 preparation used, including those due to cruciform DNA structures.
- 1102 2. A binomial filtering strategy was used to remove variants with aggregated count
1103 distributions consistent with germline single nucleotide polymorphisms.
- 1104 3. A beta-binomial filter was used to remove low-frequency artefacts, i.e. variants present
1105 at low frequencies across samples in a way not consistent with the sample-to-sample
1106 variation expected for acquired somatic mutations.
- 1107 4. Sites with a mean depth below 8 and over 40 were removed.
- 1108 5. Thresholds were used to filter out in vitro variants from the remaining mutations using
1109 a bespoke script. These were set to require a minimum variant read count of 2 or more
1110 and a variant allele fraction of 0.2 for autosomes and 0.4 for XY chromosomes.
- 1111 6. The final filtering step involved building a phylogenetic tree from the HSC genomes
1112 derived from each individual. Mutations that did not fit the optimal tree structure were
1113 also discarded as likely artefacts.

1114

1115 Tree building was performed using MPBoot, which is a maximum parsimony tree
1116 approximation method⁵¹. Variants were genotyped as 'present' in a sample if 2 or more variant
1117 reads supported the variant. Variants were genotyped as 'absent' in a sample if 0 variant reads
1118 were present at a given site and depth at that site was 6 or more. Sites that did not fall into
1119 either of the above categories were marked as 'unknown'. Mutations were assigned back to the
1120 tree using an R package (*tree_mut*), which uses a maximum likelihood approach and the
1121 original count data to assign each mutation to a branch in the MPBoot generated tree.

1122

1123 **Estimation of mutation burden in standard sequencing data**

1124

1125 Using clonal or nearly-clonal samples, we were able to compare NanoSeq to mutation burden
1126 estimates from standard whole-genome sequencing. This includes libraries prepared by laser
1127 microdissection and low-input enzymatic fragmentation³⁸ or sonication, followed by standard

1128 Illumina sequencing and mutation calling using CaVEMan⁴⁹. The mutation calls described in
1129 the previous section were further processed to make burden estimates comparable across
1130 protocols.

1131

1132 To compare NanoSeq burdens to those from standard libraries, we restricted the analysis to
1133 regions of the genome covered by at least 20 reads in the standard libraries, to minimise the
1134 impact of low coverage on mutation calling sensitivity. We also excluded the fraction of the
1135 genome flagged as *non-analysed* by CaVEMan. Given the thorough filtering strategies applied
1136 for NanoSeq, we further restricted the analysed genome to include only sites callable in
1137 NanoSeq. Finally, given that trinucleotide frequencies in the callable genome of standard
1138 libraries differ from the background genomic frequencies, burden estimates were corrected
1139 accordingly. The difference in trinucleotide frequencies was mainly due to extensive filtering
1140 of common SNPs (frequent at CpG) and the partial depletion of trinucleotides overlapping the
1141 restriction site (TGCA). Remarkably, we found that estimates of mutation burden increased by
1142 ~20% in standard sequencing data when applying these corrections, largely due to reducing the
1143 impact of low sensitivity in certain genomic regions, either due to low coverage or mapping
1144 quality problems (**Extended Data Fig 5a,b**). More details are provided in **Supplementary**
1145 **Note 7**.

1146

1147 **Bootstrapped cosine similarity**

1148

1149 Cosine similarities are frequently used to compare mutational profiles, although they do not
1150 take into account the noise introduced by the number of mutations available. Small sample
1151 sizes can cause large cosine similarity deviations from their original spectrum. If a query profile
1152 (e.g. NanoSeq result) with n mutations is to be compared to a reference profile, we can estimate
1153 the impact of small sample sizes by bootstrapping. From the reference profile we obtain 1,000
1154 random samples with size n , and then compare each of these samples back to the reference
1155 profile. We can then calculate the cosine similarities between the query and the reference
1156 profiles and compare it to the 95% interval of cosine similarities observed in the bootstrapped
1157 samples.

1158

1159 **Mutational signature analysis**

1160

1161 Mutational signatures of single-base substitutions in their trinucleotide sequence context were
1162 inferred from sets of somatic mutation counts using the sigfit (v2.0) package for R⁵². *De novo*
1163 signature extraction was performed for a range of numbers of signatures ($N = 2, \dots, 8$), using
1164 counts of mutations grouped per tissue type (cord blood, adult blood, granulocytes, colonic
1165 crypts, smooth muscle or neurons), and sequencing method (NanoSeq or standard sequencing).
1166 To account for differences in sequence composition across samples, NanoSeq mutation counts
1167 were corrected as described in a previous section (see **Correction of mutation burden and**
1168 **trinucleotide substitution profiles**). To avoid an excessive influence of tissue types more
1169 highly represented in our dataset, mutation counts were randomly downsampled to a maximum
1170 of 2,000 mutations from each tissue type. Samples with evidence of sporadic mutational
1171 processes, such as APOBEC or colibactin were removed from the dataset. This excluded
1172 urothelium, a bladder tumour sample and colonic crypts from one donor affected by colibactin
1173 (PD37449, $n = 3$). The best-supported number of signatures on the basis of overall goodness-
1174 of-fit, as reported by the 'extract_signatures' function in sigfit, was $N = 3$. The three extracted
1175 signatures (**Fig. 3e**) were subsequently fitted to the counts of mutations per sample (using the
1176 'fit_signatures' function in sigfit) to infer the exposure of each signature in each sample.

1177

1178 Mutational signature analysis was also applied to publicly-available single-nuclei mutation
1179 data from neurons¹³. Three signatures closely matching those shown in the original publication
1180 were extracted using the *extract_signatures* function in sigfit, with parameters nsignatures=3,
1181 seed=1469 and iter=10000.

1182

1183 **Linear regression modelling**

1184

1185 Linear regressions were used to estimate the numbers of mutations accumulated per year, to
1186 test whether mutations associated with a given signature increased with age, or to test the
1187 effects of disease status or organ of origin on mutation burdens.

1188

1189 In analyses where only one sample per donor was available, we used multiple linear regression.
1190 For **Fig. 3o,p**, which show the mutation rate per year across tissues, we used linear regressions
1191 without an intercept, as all tissues had intercepts close to and not-significantly different from
1192 zero.

1193

1194 For those cases with multiple samples per donor, including smooth muscle, colonic crypts,
1195 granulocytes or sperm, we used linear mixed-effects models, using donor as a random effect
1196 (random slopes). For example, using formulas such as: mutations ~ age + (age - 1|donor). This
1197 enabled us to account for the relatedness of multiple samples per donor.

1198

1199 To test for the significance of a given fixed effect (such as organ of origin), we used Likelihood
1200 Ratio Tests using the *anova* R function, comparing the null model without the fixed effect and
1201 the alternative model with the fixed effect. Confidence intervals for linear mixed-effects
1202 models were calculated using parametric bootstrapping and 1,000 replicates, as implemented in
1203 the 'predict' method in bootpredictlme4 R package.

1204

1205 All linear regression and statistical tests were conducted in R using packages: *lm*, *lmer*, *afex*,
1206 *bootpredictlme4*, and *lmerTest*.

1207

1208

1209 **Supplementary References**

1210

1211 38 Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture
1212 microdissection and low-input DNA sequencing. *Nat Protoc*, doi:10.1038/s41596-
1213 020-00437-6 (2020).

1214 39 Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell*
1215 **182**, 672-684.e611, doi:10.1016/j.cell.2020.06.036 (2020).

1216 40 Krishnaswami, S. R. *et al.* Using single nuclei for RNA-seq to capture the
1217 transcriptome of postmortem neurons. *Nat Protoc* **11**, 499-524,
1218 doi:10.1038/nprot.2016.015 (2016).

1219 41 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
1220 MEM. *arXiv: Genomics* (2013).

1221 42 Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms
1222 on BAM files. *Source Code Biol Med* **9**, 13-13, doi:10.1186/1751-0473-9-13 (2014).

1223 43 Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in
1224 tumour cell populations. *Nat Commun* **3**, 811, doi:10.1038/ncomms1814 (2012).

1225 44 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation
1226 in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

1227 45 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74,
1228 doi:10.1038/nature15393 (2015).

1229 46 Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from
1230 sequence reads. *Genome Res* **30**, 185-194, doi:10.1101/gr.246934.118 (2020).

1231 47 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in
1232 high-throughput sequencing. *Nucleic Acids Res* **40**, e72, doi:10.1093/nar/gks001
1233 (2012).

1234 48 Robinson, P. S. *et al.* Elevated somatic mutation burdens in normal human cells due
1235 to defective DNA polymerases. *bioRxiv*, 2020.2006.2023.167668,
1236 doi:10.1101/2020.06.23.167668 (2020).

1237 49 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to
1238 Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics*
1239 **56**, 15.10.11-15.10.18, doi:10.1002/cpbi.20 (2016).

1240 50 Raine, K. M. *et al.* cgpPindel: Identifying Somatic Acquired Insertion and
1241 Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**,
1242 15.17.11-15.17.12, doi:10.1002/0471250953.bi1507s52 (2015).

1243 51 Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference
1244 and bootstrap approximation. *BMC Evol Biol* **18**, 11, doi:10.1186/s12862-018-1131-3
1245 (2018).

1246 52 Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational
1247 signatures. *bioRxiv*, 372896, doi:10.1101/372896 (2020).

1248 53 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*
1249 **518**, 317-330, doi:10.1038/nature14248 (2015).

1250 54 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation
1251 sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513,
1252 doi:10.1073/pnas.1208715109 (2012).

1253 55 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the
1254 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**,
1255 11.10.11-11.10.33, doi:10.1002/0471250953.bi1110s43 (2013).

1256 56 Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to
1257 benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623-630,
1258 doi:10.1038/nmeth.3407 (2015).

1259 57 Catlin, S. N., Busque, L., Gale, R. E., Gutter, P. & Abkowitz, J. L. The replication
1260 rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460-4466,
1261 doi:10.1182/blood-2010-08-303537 (2011).

1262 58 Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex
1263 differentiation landscapes. *Nature* **553**, 418-426, doi:10.1038/nature25022 (2018).

1264 59 Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal
1265 hematopoiesis. *Science* **367**, 1449-1454, doi:10.1126/science.aay9333 (2020).

1266 60 Summers, C. *et al.* Neutrophil kinetics in health and disease. *Trends Immunol* **31**,
1267 318-324, doi:10.1016/j.it.2010.05.006 (2010).

1268 61 Abkowitz, J. L., Catlin, S. N. & Gutter, P. Evidence that hematopoiesis may be a
1269 stochastic process in vivo. *Nat Med* **2**, 190-197, doi:10.1038/nm0296-190 (1996).

1270 62 Abkowitz, J. L., Golinelli, D., Harrison, D. E. & Gutter, P. In vivo kinetics of
1271 murine hemopoietic stem cells. *Blood* **96**, 3399-3405 (2000).

1272 63 Derényi, I. & Szöllösi, G. J. Hierarchical tissue organization as a general mechanism
1273 to limit the accumulation of somatic mutations. *Nat Commun* **8**, 14545,
1274 doi:10.1038/ncomms14545 (2017).

1275

1276 **Data Availability**

1277

1278 Information on data availability for all samples is available in **Supplementary Table 1**.
1279 NanoSeq sequencing data has been deposited in EGA under accession number
1280 EGAD00001006459. Standard sequencing data has been deposited in EGA under accession
1281 number [pending submission]. For samples publicly available, references to the original
1282 sources are provided in **Supplementary Table 1**. Substitution and indel calls for samples
1283 sequenced with NanoSeq are available in **Supplementary Tables 4 and 5**.

1284

1285 **Code Availability**

1286

1287 The bioinformatic pipeline to process NanoSeq sequencing data includes all steps from
1288 processing sequencing data, mapping, calling mutations and calculating corrected burden
1289 estimates and substitution profiles. This code is available from
1290 <https://github.com/cancerit/NanoSeq>. Pipelines to call indels, do signature extraction and
1291 signature fitting with SigFit, simulate efficiency of the NanoSeq protocol, and to calculate
1292 mutation burden in specific chromosomal regions, are available from
1293 https://github.com/fa8sanger/NanoSeq_Paper_Code.

1294

1295 **Acknowledgements**

1296

1297 We thank Liz Anderson, Kirsty Roberts, Calli Latimer, Quan Lin, the CGP-lab, Rocio Vicario,
1298 Frederic Geissmann, Nicos Angelopoulos, German Tischler, Tristram Bellerby, Maria Abascal
1299 and Krishnaa Chatterjee for assistance in the development of NanoSeq or with this manuscript.

1300

1301 We are grateful to the live donors and the families of the deceased transplant organ donors.
1302 This research was supported by the Cambridge NIHR BRC Cell Phenotyping Hub. We
1303 gratefully acknowledge the participation of all NIHR BioResource Centre
1304 Cambridge volunteers, and thank the NIHR BioResource Centre Cambridge and staff for their
1305 contribution. We thank the National Institute for Health Research and NHS Blood and
1306 Transplant. The views expressed are those of the author(s) and not necessarily those of the
1307 NHS, the NIHR or the Department of Health & Social Care. We gratefully acknowledge the
1308 Cambridge Blood and Stem Cell Biobank for sample donation and support of this work. We
1309 are grateful to the Cambridge Brain Bank for sample donation.

1310

1311 **Funding:** I.M. is funded by Cancer Research UK (C57387/A21777) and the Wellcome Trust.
1312 P.J.C. is a Wellcome Trust Senior Clinical Fellow. R.R. is a recipient of a CRUK Career
1313 Development fellowship (C66259/A27114). E.L. is supported by a Wellcome/Royal Society
1314 Sir Henry Dale Fellowship (Grant number 107630/Z/15/Z), the European Hematology
1315 Association, BBSRC and by core funding from Wellcome (Grant number 203151/Z/16/Z) and
1316 MRC to the Wellcome-MRC Cambridge Stem Cell Institute. D.G.K. is supported by a
1317 Bloodwise Bennett Fellowship (15008), the Bill and Melinda Gates Foundation (INV-002189)
1318 and an ERC Starting Grant (ERC-2016-STG-715371).

1319

1320 **Author Contributions**

1321

1322 R.J.O., F.A., and I.M. conceived the project. I.M., P.J.C., R.R., and M.R.S. supervised the
1323 project. F.A., R.J.O., E.M., and I.M. wrote the manuscript; all authors reviewed and edited the
1324 manuscript. R.J.O. led the development of the protocol with help from F.A., A.R.J.L., P.E.,

1325 S.V.L. and I.M. R.J.O. and F.A. developed the bioinformatics pipeline with help from R.E.A.,
1326 S.V.L., and D.J. F.A. led the analysis of the data with help from A.R.J.L., I.M., A.B-O., Y.W.,
1327 L.M.R.H., E.J.K., T.H.H.C, M.S.C, and M.G. E.M. performed the HSC/MPP experiments.
1328 L.M.R.H. and A.J.C.R. performed the cell sorting of neuronal nuclei. A.R.J.L. and A.C.
1329 performed laser microdissection. E.M., N.F.O., H.E.M., M.D., D.G.K., E.L., K.T.A.M., K.S.P.,
1330 K.A., R.R., H.L.S. and S.O collected and processed samples. E.M., E.L., M.G. and D.G.K
1331 assisted on the interpretation of blood data.

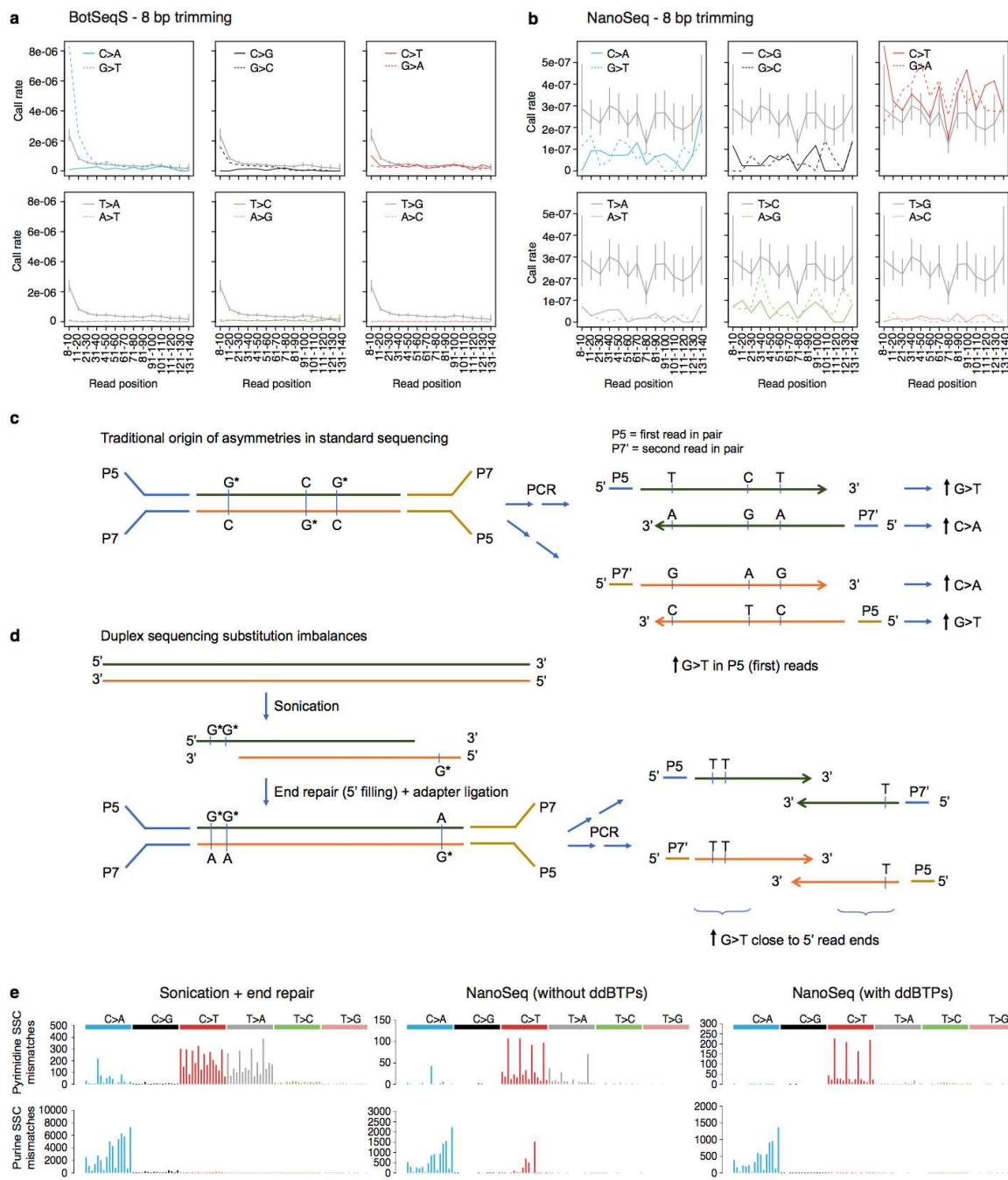
1332

1333 **Competing Interests Declaration**

1334

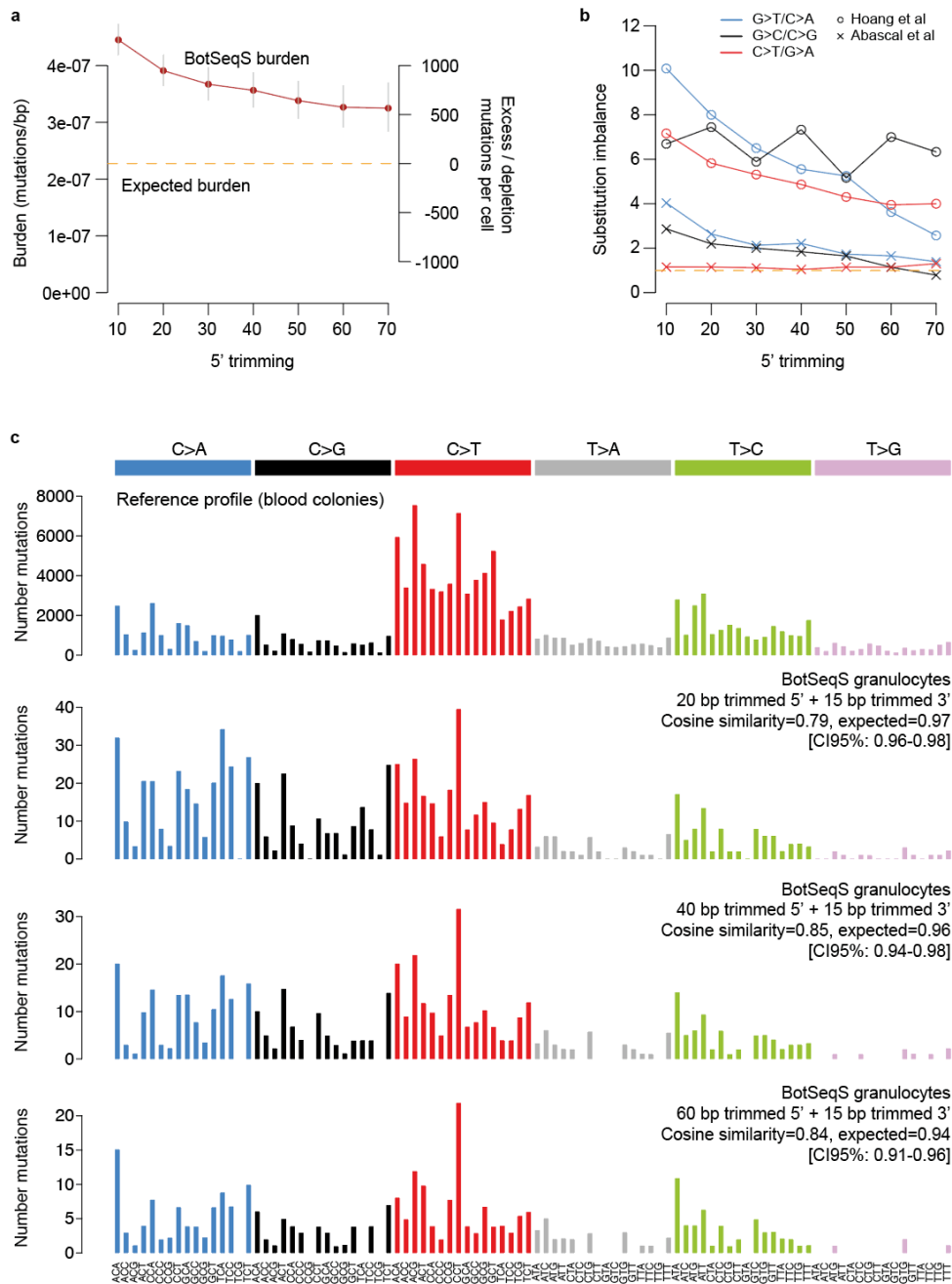
1335 A patent application on NanoSeq is being filed including several authors.

1336



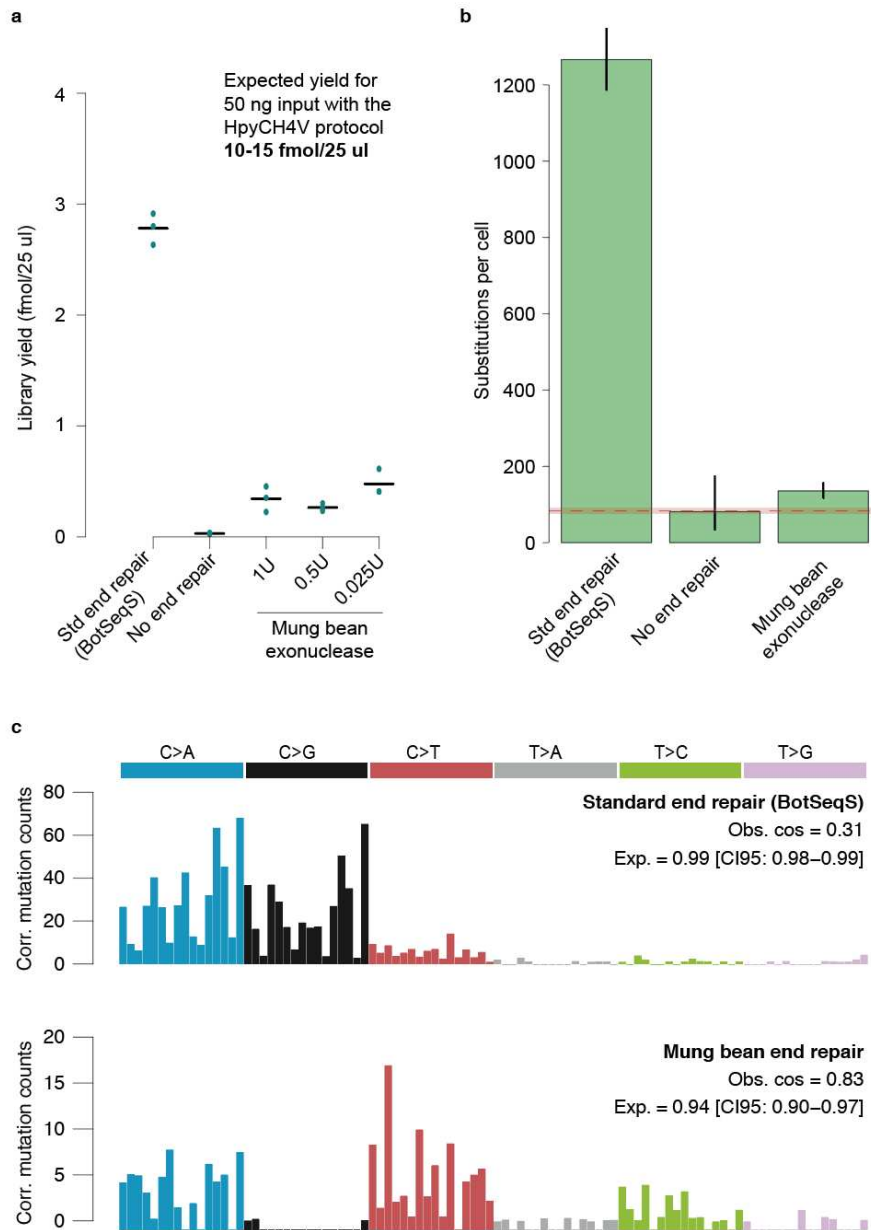
1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349

Extended Data Figure 1 | Substitution imbalances and impact of A-tailing. **a-b**, Imbalances in the distribution of the six complementary substitutions (e.g. G>T vs C>A) across read positions in BotSeqS and NanoSeq, respectively. **c**, Origin of G>T over C>A mutation call imbalances in standard sequencing²². **d**, Origin of imbalances in Duplex Sequencing / BotSeqS as a result of end repair during library preparation. **e**, Single-strand consensus calls for pyrimidine (top) and purine (bottom) substitutions for the standard BotSeqS (left panel) protocol and for NanoSeq with standard and modified A-tailing protocols (middle and right panels, respectively). For example, C>T changes are shown on the top, while the complementary G>A changes are shown on bottom. By using ddBTPs C>A, G>A and T>A errors are reduced, lowering the risk of false positive double-strand consensus calls.



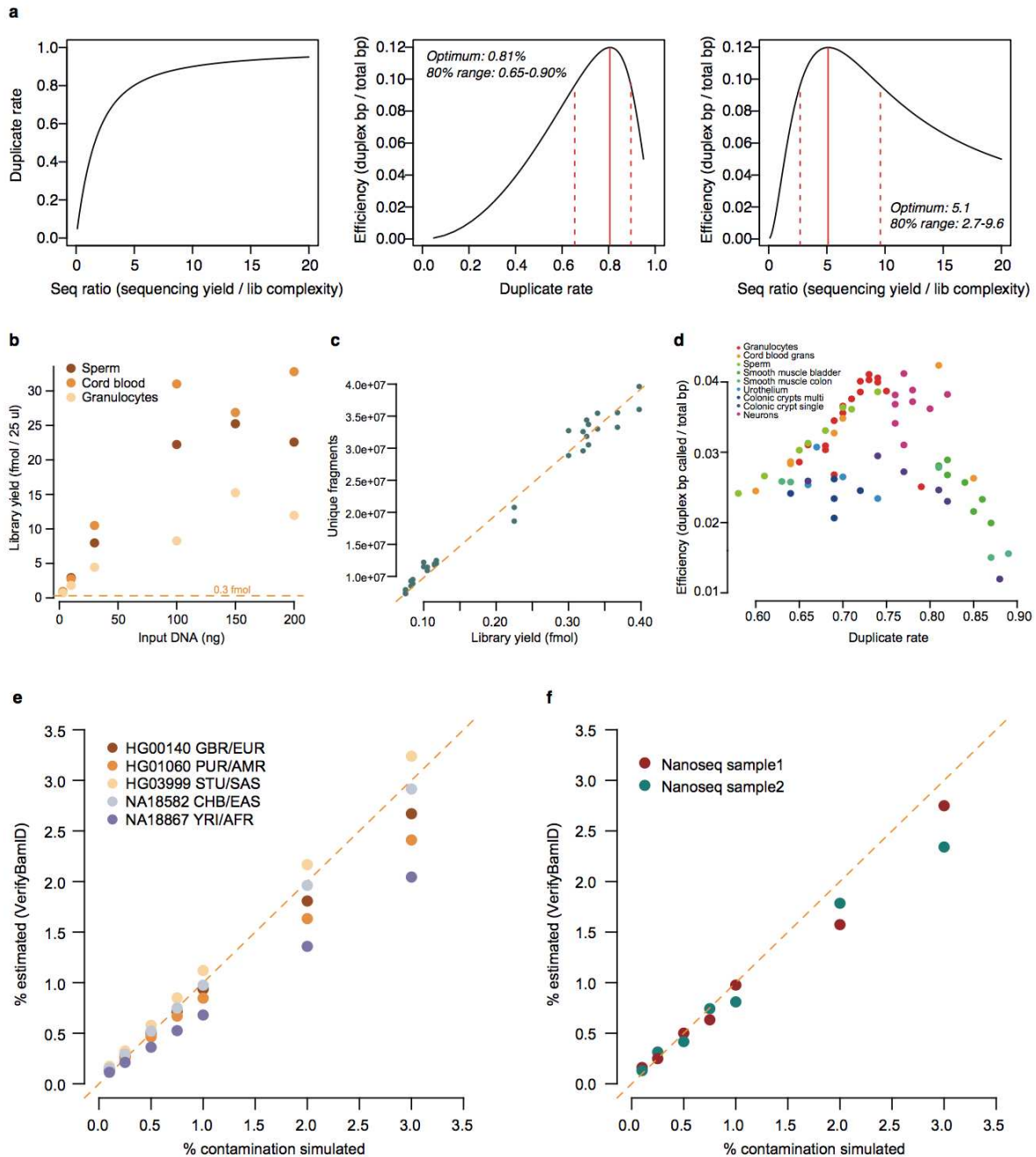
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363

Extended Data Figure 2 | BotSeqS errors as a function of read end trimming. **a**, BotSeqS estimated burden for the granulocyte sample shown in **Fig 1b-d** applying different trimmings to the 5' ends of reads. Even with extensive trimming we predict at least ~600 artefactual mutation calls per diploid genome. **b**, Substitution imbalances are observed deep into the reads and cannot be avoided with read trimming. Imbalances vary from experiment to experiment, as a consequence of DNA damage on the DNA source or during library preparation (**Supplementary Note 1**). **c**, Substitution profiles including the reference profile from single-cell derived blood colonies and three BotSeqS profiles after trimming of 20, 40 and 60 bp from the 5' end of reads (in addition to 15 bp trimming of the 3' end). The text in the figure indicates the observed and expected cosine similarities (**Methods**) cosine similarity to the reference profile. C>A and C>G errors in BotSeqS remain after extensive trimming.



1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379

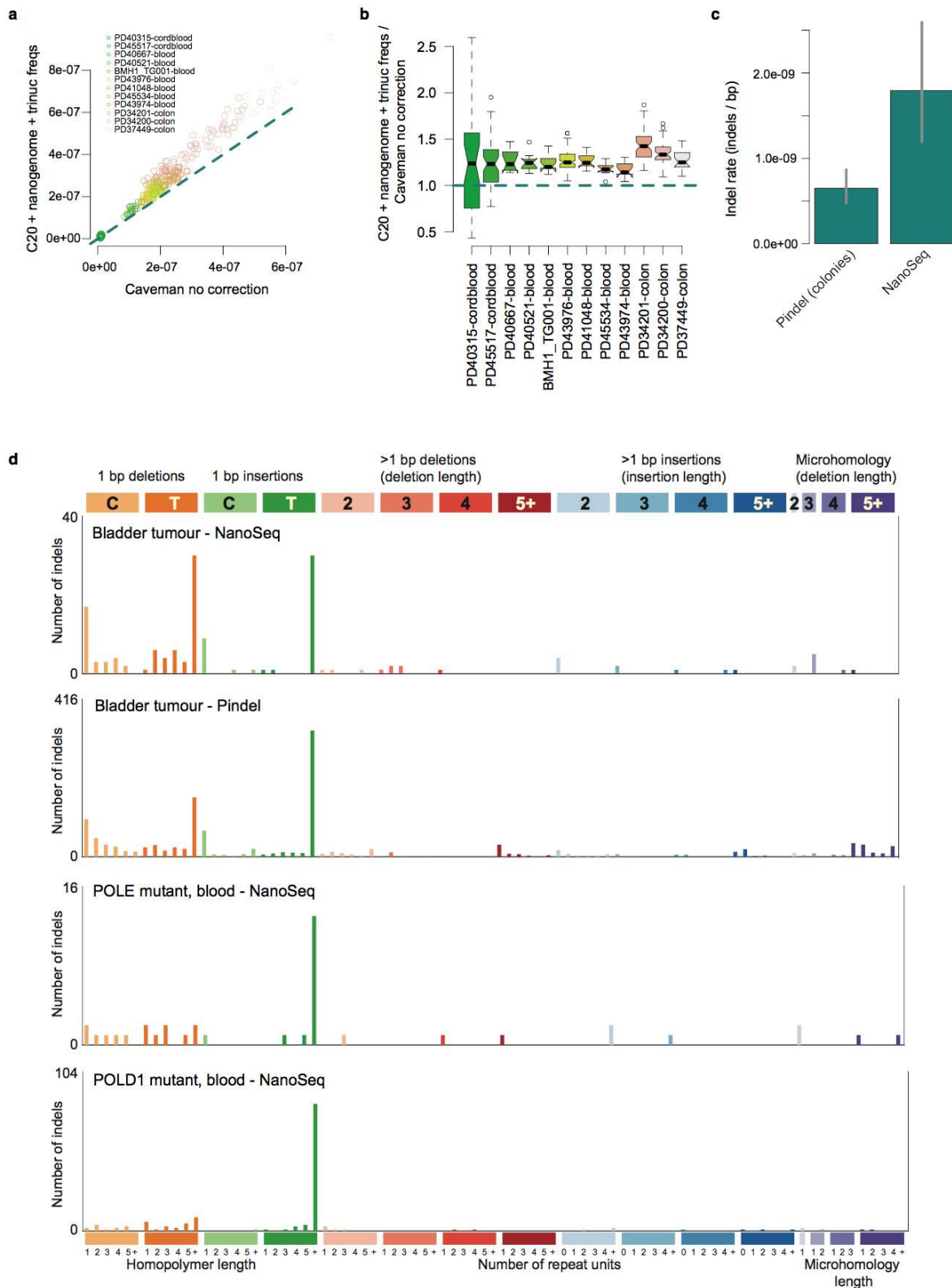
Extended Data Figure 3 | Alternative protocols for library preparation. **a**, Library preparation yields for three different kind of protocols run in triplicates (green dots show replicates; mean values as black lines). For Mung bean, different concentrations of the enzyme (U) were tested. **b**, Estimated number of mutations per cord blood cell. Poisson 95% confidence intervals are shown as lines. The red dotted line shows the number of mutations per cord blood cell estimated with the restriction enzyme NanoSeq protocol, with Poisson 95% confidence intervals shown as a red shade. In contrast to **Fig 1g**, we did not apply the correction for missing embryonic mutations because here we are comparing three protocols that are equally affected by this limitation. **c**, Substitution profiles for the standard end repair protocol (BotSeqS) and for Mung Bean, showing the cosine similarities with the reference profile (**Fig 1h**). The profile for the protocol without any end repair is not shown because the very low library preparation yields limited the detection of mutations.



1380
1381

1382 **Extended Data Figure 4 | Optimization of duplicate rates, DNA input requirements and**
 1383 **estimation of human contamination.** **a**, Relationship between sequencing yield, library
 1384 complexity, duplicate rates and efficiency, based on a truncated Poisson model (**Methods**).
 1385 From left to right: duplicate rate as a function of the sequencing ratio (sequencing reads / DNA
 1386 fragments in the library); efficiency (measured as bases called with duplex coverage/bases
 1387 sequenced) as a function of the duplicate rate; and efficiency as a function of sequencing ratio.
 1388 **b**, Library yield as fmol per 25 μ l as a function of the amount of input DNA in ng. **c**, Empirical
 1389 relationship between the estimated fmol in library (measured by qPCR) and the number of
 1390 unique molecules in the library estimated with Picard tools (Lander-Waterman equation) for
 1391 our choice of restriction enzyme and fragment size selection (250 - 500 bp). **d**, Empirical
 1392 relationship between duplicate rates and efficiency of the method, measured as duplex bases
 1393 called / number of bases sequenced (i.e. the number of paired-end reads multiplied by 300).

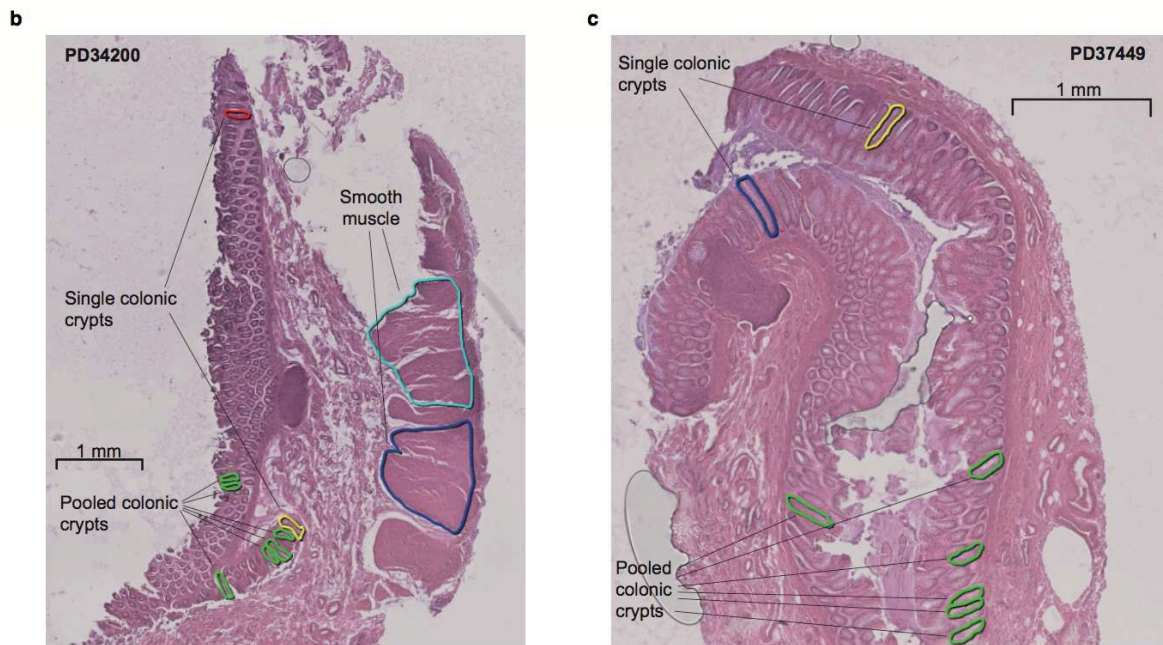
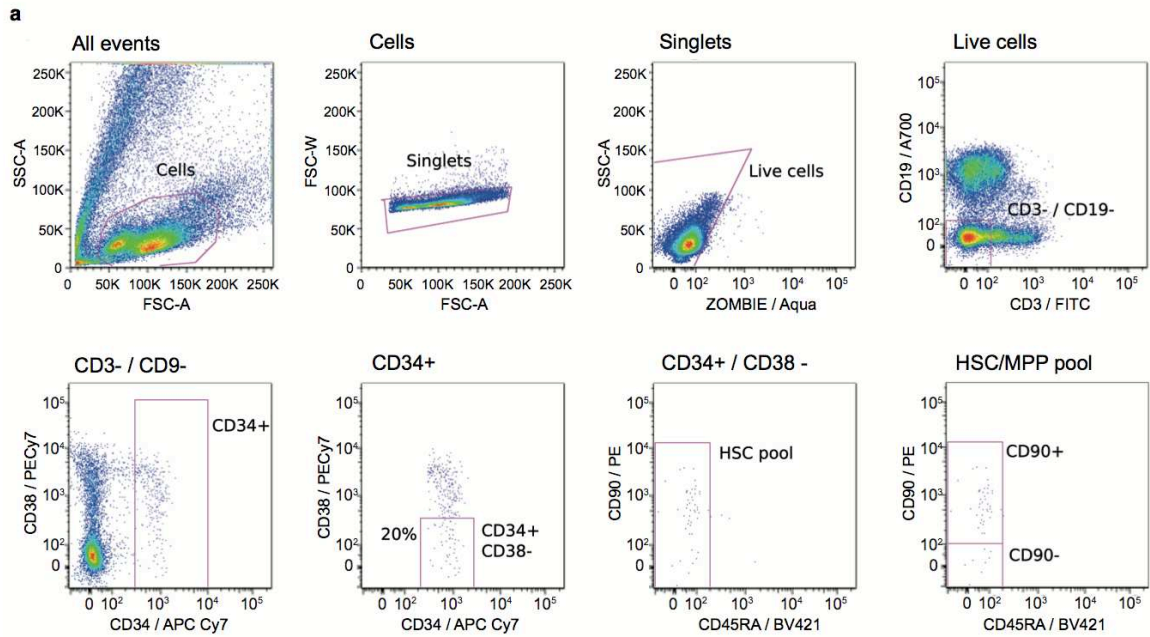
1394 The maximum efficiency (~ 0.04) is lower than the maximum analytical expectation (0.12;
1395 middle panel in **(a)** because of the trimming of read ends (barcodes, restriction sites and 8 bps
1396 from each end) and the strict filters that we apply to consider a site callable. **e**, VerifyBamId
1397 contamination estimates for different amounts of simulated contamination from individuals of
1398 different ancestry. **f**, Contamination simulation using two NanoSeq samples to contaminate
1399 each other.



1400
 1401
 1402
 1403
 1404
 1405
 1406
 1407

Extended Data Figure 5 | Correction of standard (CaVEMan-based) mutation burden estimates and validation of NanoSeq indel. **a**, Comparison of the mutation burden estimates in regions of the genome with at least 20x coverage (*c*) to the trinucleotide-context-corrected mutation burdens in the subset of *c* covered by NanoSeq and passing all NanoSeq filters. **b**, Ratio between the rates shown in panel (*a*), showing that the corrected burden is approximately 20% higher than the uncorrected burden. **c**, Comparison of indel rates between cord blood

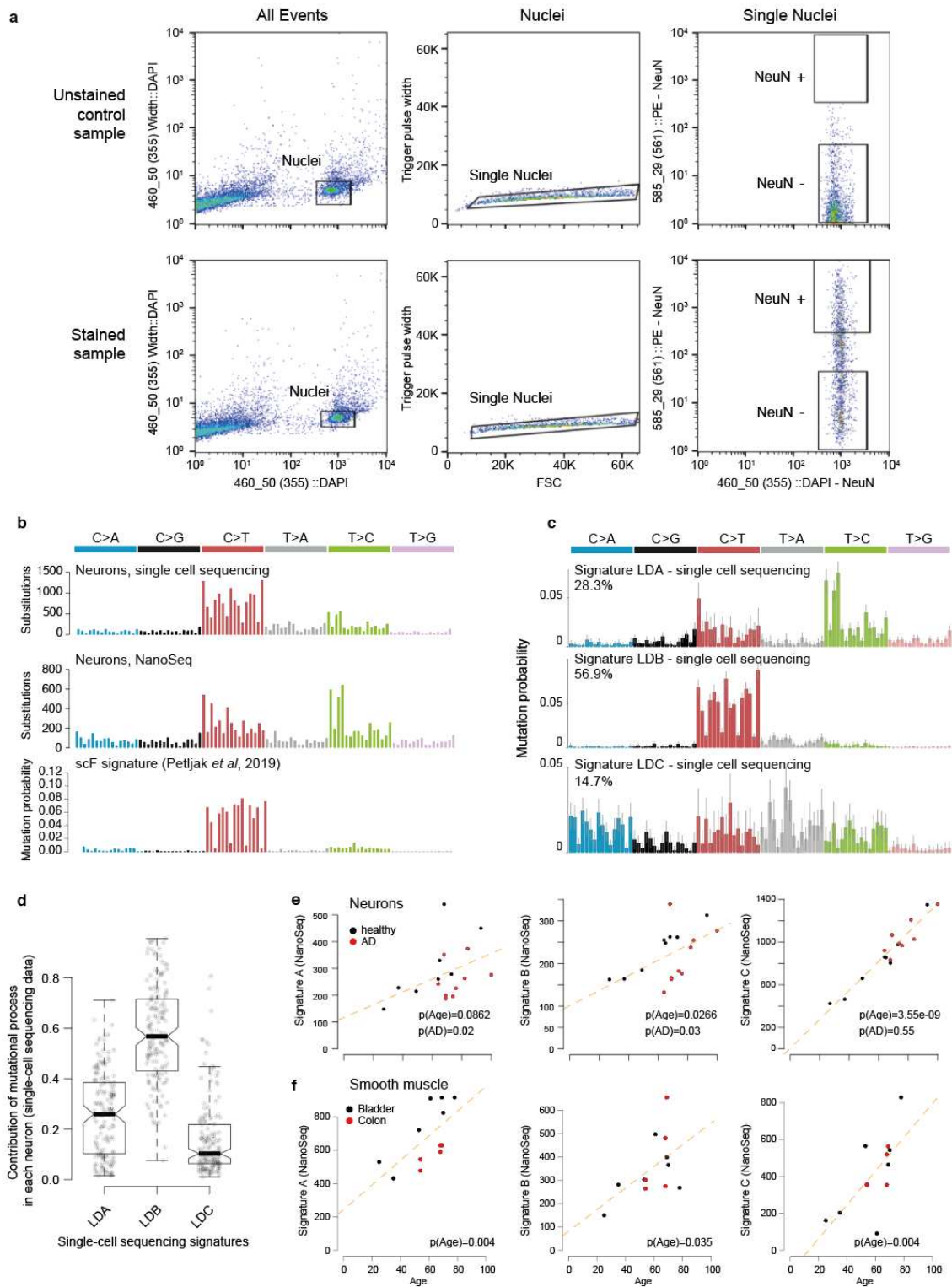
1408 colonies (indels were called with the Pindel algorithm) and granulocytes from neonates
1409 (NanoSeq pipeline). **d**, The top two panels show the high similarity between the NanoSeq and
1410 Pindel indel profiles for a bladder tumour; the bottom two profiles show the indel spectra in
1411 blood from *POLE* and a *POLD1* germline mutation carriers, very similar to the reported
1412 profiles in Robinson *et al*⁴⁸.



1413
1414

1415 **Extended Data Figure 6 | Haematopoietic stem and progenitor cells and colon histology.**

1416 **a**, Gating strategy for the isolation of HSC/MPPs from a representative bone marrow sample.
 1417 Text above plots indicates the population depicted. Text inside the plots indicates the name of
 1418 the gates shown in pink. The CD34+/CD38- population is defined as the bottom 20% CD38-
 1419 as shown. For all initial samples (BM/PB/CB) the index sorted population is the "HSC pool"
 1420 gate. Cell population abundance differed between samples but typically viable cells were 60-
 1421 90% of total cells and singlets were 98-99% of viable cells. Live cells were 90-99% of viable
 1422 cells and myeloid cells were 15-50% of live cells. CD34+ cells were typically 1-15% of
 1423 myeloid cells. **b** and **c**, Colon histology sections showing microbiopsied areas of colonic
 1424 epithelium and smooth muscle for donors PD34200 and PD37449, respectively.
 1425



1426

1427

1428 **Extended Data Figure 7 | Neuron nuclei sorting, comparison to single-cell data and**

1429 **accumulation of mutations with age.** **a**, Gating strategy for the isolation of neuronal nuclei

1430 from frontal cortex. Nuclei were sorted by FACS using an Influx cell sorter (BD Biosciences)

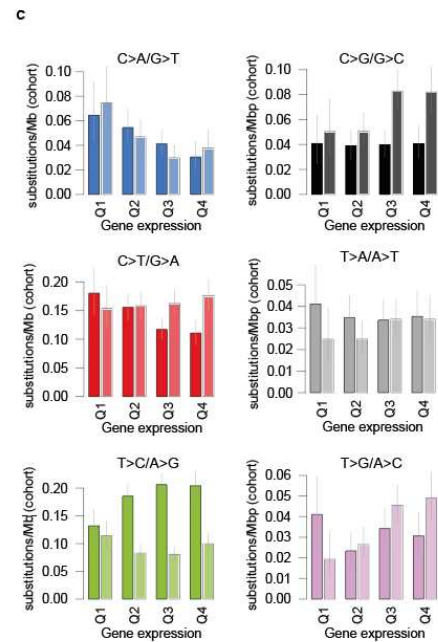
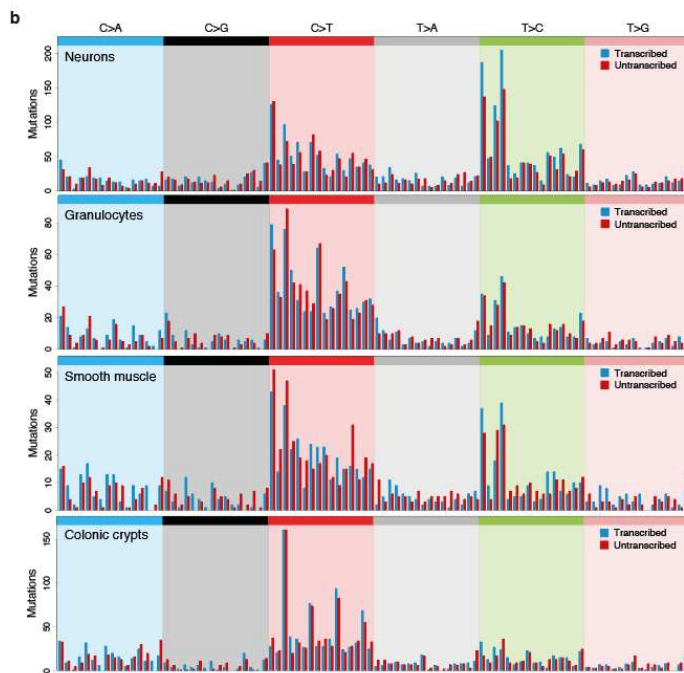
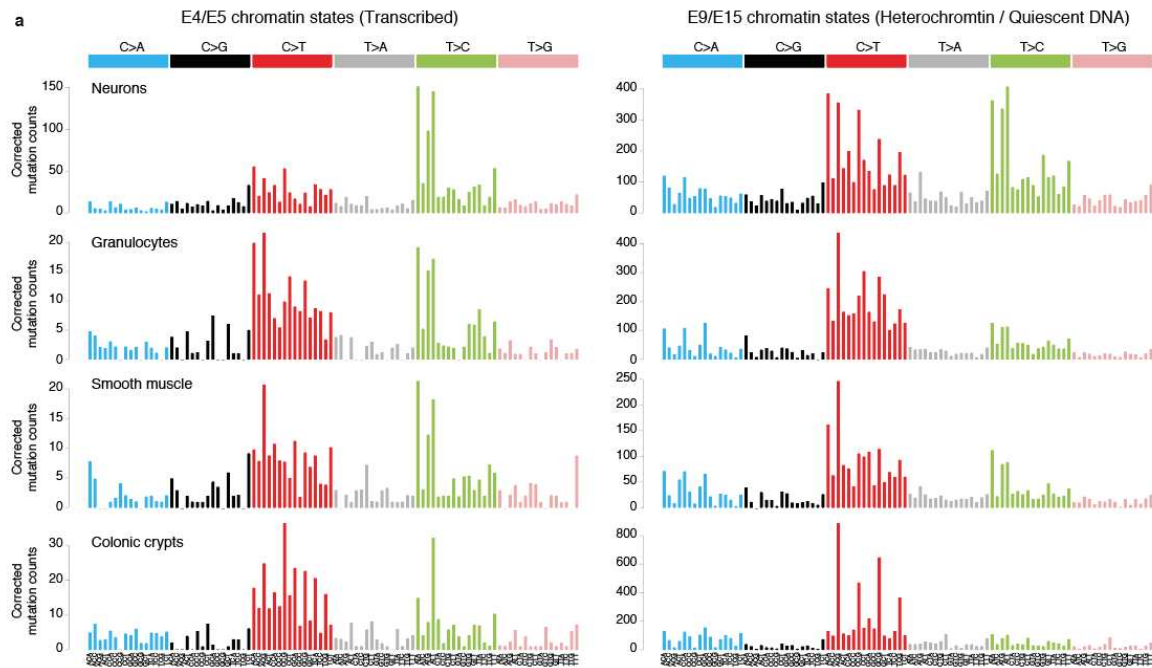
1431 with a 100- μm nozzle. For each sample an unstained control was used to help determine the

1432 NeuN+ population. The text above each column indicates the population depicted and the text

1433 inside the plots indicates the population of the gates highlighted in black. Sorting results varied

1434 among samples, with 1-60% passing the DAPI gate and, of these, 2-53% passing a conservative

1435 NeuN+ gate. **b**, Substitution profiles for all mutations detected in neurons with SNP-phased
1436 error-corrected single-cell sequencing data in Lodato *et al.*¹³ (top) and with NanoSeq (middle).
1437 In the bottom panel, a signature specific of single-cell sequencing data is shown (scF signature
1438 from Petjak *et al.*¹⁵). **c**, Mutational signatures extracted from Lodato *et al.*¹³, showing their
1439 relative contributions in the published dataset. These signatures were obtained using SigFit
1440 (**Methods**) on publicly-available mutation calls and are referred to as LDA, LDB and LDC.
1441 Note the high similarity between the NanoSeq full spectrum for neurons and LDA (cosine
1442 similarity 0.96), and between scF and LDB (cosine similarity 0.97). **d**, Predicted contribution
1443 of LDA, LDB and LDC to each of the neurons sequenced in Lodato *et al.*¹³. **e**, Accumulation
1444 of mutations attributed to NanoSeq signatures A, B, and C with age in healthy donors and in
1445 Alzheimer's disease donors. **f**, Accumulation of mutations attributed to NanoSeq signatures A,
1446 B, and C in smooth muscle from bladder and colon.



1447
1448

1449 **Extended Data Figure 8 | Normalised substitution spectra across different genomic**

1450 **regions. a**, Substitution spectra for neurons, granulocytes, smooth muscle and colonic crypts

1451 in chromatin states associated to transcription (states E4 and E5 in ENCODE), and inactive

1452 DNA (E9 and E15). Chromatin states were obtained from ENCODE⁵³, using the following

1453 epigenomes: E073 (frontal cortex), E030 (granulocytes), E076 (smooth muscle), and E075

1454 (colonic mucosa). To enable direct comparison of spectra across genomic regions with

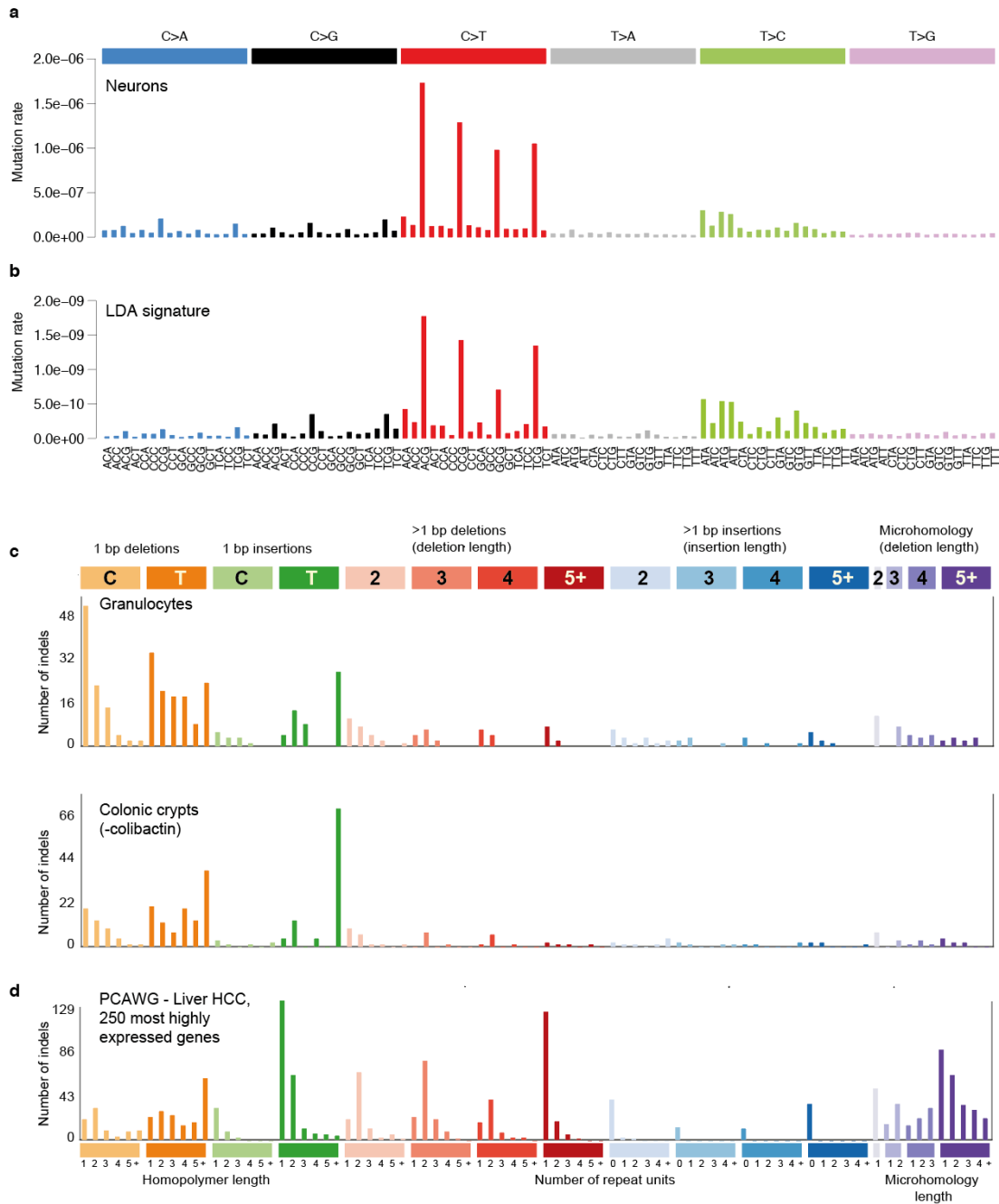
1455 different trinucleotide frequencies, the profiles have been normalised to the genomic

1456 trinucleotide frequencies (**Methods**). **b**, Transcriptional strand asymmetries in neurons,

1457 granulocytes, smooth muscle and colonic crypts. **c**, Transcriptional strand asymmetries in

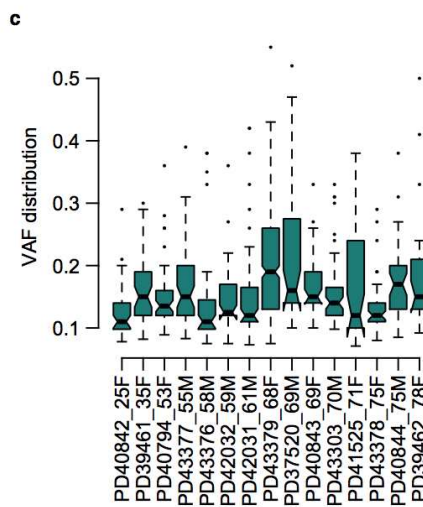
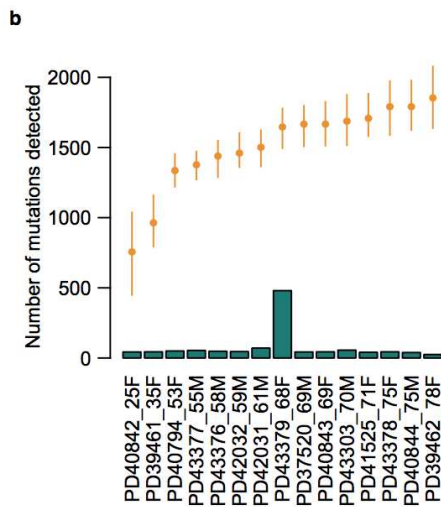
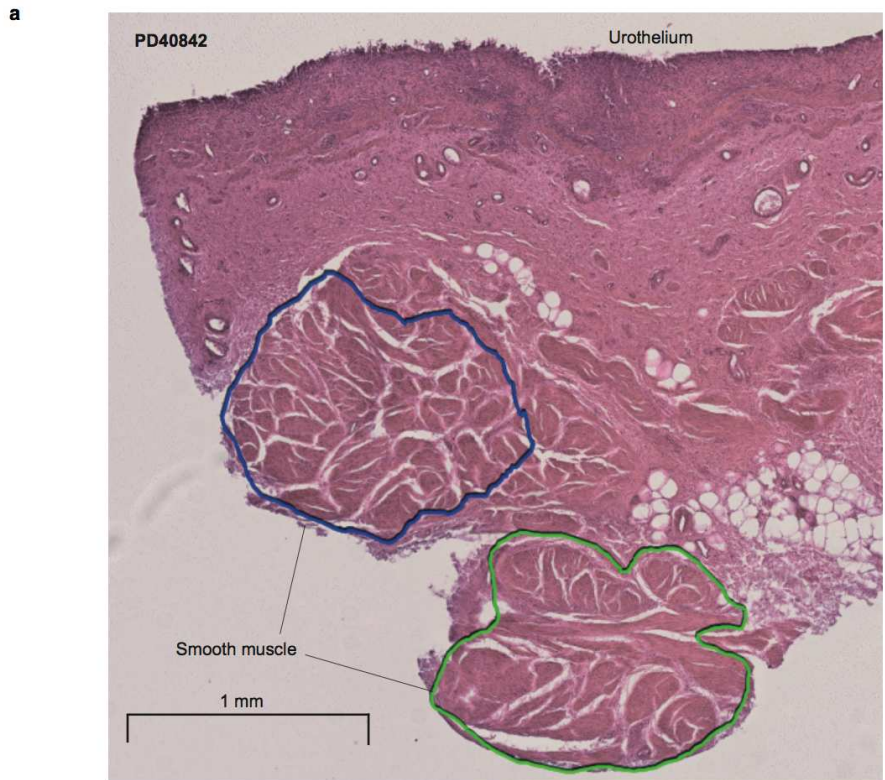
1458 neurons in quartiles of gene expression.

1459



1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472

Extended Data Figure 9 | Additional substitution and indel spectra. **a**, NanoSeq mutational spectrum for neurons corrected for trinucleotide frequency in the callable genome. Unlike the usual representation, which shows unnormalized rates, this representation shows mutation rates per available trinucleotide. **b**, LDA signature from Lodato *et al.*¹³ normalised for trinucleotide frequency in the genome also reveals high C>T rates at CpG dinucleotides. This observation from single-cell data suggests that the high C>T rates at CpG sites in NanoSeq neuron data (**a**) is not caused by contamination of NeuN+ pools with glia or other cells. **c**, Indel profiles of granulocytes (top) and of colonic crypts without the colibactin signature (bottom). **d**, Indel profiles for the 250 most highly expressed genes in PCAWG liver hepatocellular carcinoma data³¹.



1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483

Extended Data Figure 10 | Smooth muscle. **a**, Histology of bladder smooth muscle showing two sections from donor PD40842. **b**, Number of mutations detected with CaVEMan in different smooth muscle sections processed with our standard microdissection sequencing protocol³⁸. The orange dots show the expected mutation burdens (with 95% confidence intervals) for these sections based on the donor age and the regression model shown in Fig. 3j. **c**, Distribution of variant allele frequencies (VAFs) for each of the smooth muscle sections using standard whole-genome sequencing. Boxplot notches show the 95% confidence interval for the median.

1484 **Supplementary Notes**

1485

1486 **Supplementary Note 1 - Single strand consensus calls**

1487

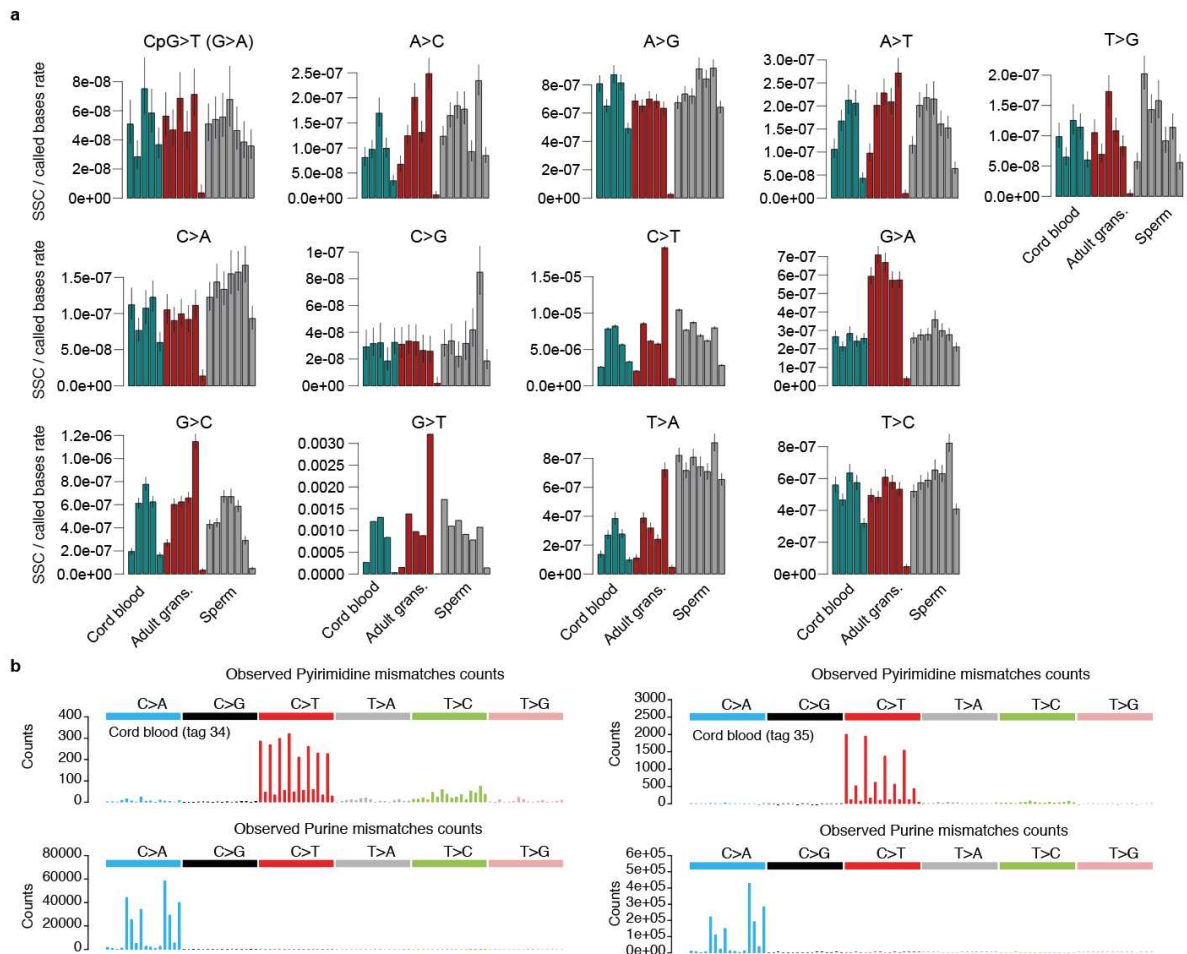
1488 In duplex sequencing protocols, consensus calls from only one of the two strands (single strand
1489 consensus calls) can be caused by amplification artefacts or by DNA damage⁵⁴, including
1490 damage originated in vivo or in vitro, during DNA extraction, storage or library preparation.
1491 Given the value of measuring in-vivo DNA damage, we explored whether information on DNA
1492 damage could be extracted from the abundant single strand consensus (SSC) calls.

1493

1494 In our data, SSC calls are dominated by G>T, G>C and C>T substitutions, with almost
1495 complete asymmetries between the rates of these changes and the complementary substitutions.
1496 To explore the extent of biological and technical variation in SSC calls, we analysed data from
1497 three different samples (cord blood, sperm and adult granulocytes) with multiple replicate
1498 NanoSeq libraries, some of which were made and sequenced at different times. This analysis
1499 revealed large variation in SSC substitution rates and spectra between technical replicates
1500 (**Supplementary figure 1a**). For example, the 6th granulocyte replicate library (shown in
1501 **Supplementary figure 1a** as the rightmost red bar) was made and sequenced on a different
1502 day to the other five, and has a markedly different SSC pattern. The 96-bar substitution spectra
1503 for two cord blood libraries from the same sample further show how the pattern of SSC calls
1504 varies between libraries from the same sample (**Supplementary figure 1b**).

1505

1506 Overall, the large variation in SSC profiles between replicates suggests that the vast majority
1507 of SSC calls in our data represent technical artefacts, likely resulting from DNA damage
1508 introduced during library preparation, rather than pre-existing DNA damage in the input DNA.



1509

1510

Supplementary Figure 1. a, Rate of single strand consensus calls in replicates of cord blood, adult granulocytes and sperm samples. **b**, Substitution profiles separated for purines and pyrimidine substitutions for two cord blood libraries from the same donor.

1511

1512

1513

1514

1515

Supplementary Note 2 - Restriction enzyme choice

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

Supplementary Note 3 – Alternative fragmentation: sonication followed by mung bean nuclease blunting

1530

1531

1532

Restriction enzymes have several useful features in the context of NanoSeq: (1) they provide clean genome fragmentation with sufficiently representative coverage of the genome to enable

1533 accurate estimation of mutation burden and signatures, (2) they enable library preparation from
1534 low inputs of DNA, including laser-microdissection of a few hundred cells from histology
1535 sections (a minimum of 1 ng of input DNA is required for the sample coverages used in this
1536 study; **Extended Data Fig 4**), and (3) their partial coverage of the genome reduces the cost of
1537 sequencing a matched normal sample (to remove germline mutations) by ~70%, by sequencing
1538 an undiluted NanoSeq library (**Methods**).

1539

1540 However, incomplete genome representation can be undesirable for other applications, such as
1541 targeted NanoSeq. We reasoned that sonication followed by exonuclease digestion of
1542 overhangs, could provide an alternative fragmentation strategy without the errors associated
1543 with filling 5' ends in standard end repair. To compare several protocols, we used cord blood
1544 granulocytes from donor S1 in **Fig. 1g** (EM_A1_XN3325). Using 50 ng of sonicated DNA per
1545 condition, we generated libraries in triplicate using: (1) standard end repair (BotSeqS), (2) no
1546 end-repair (to quantify the frequency of blunt ended fragments generated directly by
1547 sonication), and (3) three different concentrations of Mung Bean nuclease (0.025U, 0.5U and
1548 1U per reaction, NEB M0250).

1549

1550 Library yields varied modestly among replicates but greatly among conditions (**Extended Data**
1551 **Fig 3a**). Sonication followed by standard end repair produced library yields around 20-30% of
1552 those typically obtained with HpyCH4V. Sonication followed by Mung Bean nuclease
1553 produced comparable yields across the range of exonuclease concentrations tested and around
1554 2-10% of those obtained with the HpyCH4V restriction enzyme protocol. Sonication followed
1555 by A-tailing and adapter ligation, without end repair or exonuclease blunting, produced
1556 libraries with yields ~0.3% of those using restriction enzymes, yielding much less than the
1557 required 0.3 fmol used for sequencing, and resulting in low callable coverages.

1558

1559 We then compared the mutation burden and mutational spectra across protocols (**Extended**
1560 **Data Fig 3b**). As expected, sonication followed by standard end repair (BotSeqS) yielded a
1561 high error rate, with around 1,200 errors per diploid genome ($\sim 2 \times 10^{-7}$ errors/bp) and a
1562 mutational spectrum dominated by C>A and C>G errors (**Extended Data Fig 3c**). Sonication
1563 followed by Mung Bean nuclease or no end repair yielded low mutation burdens, similar to
1564 those using the restriction enzyme protocol, with error rates estimated to be in the nano scale
1565 ($< 10^{-8}$ errors/bp). Libraries generated without end repair or Mung Bean nuclease did not
1566 produce enough library yield to enable a detailed comparison of mutation burdens and spectra.
1567 The mutational spectra of the Mung Bean nuclease libraries were largely consistent with that
1568 of cord blood single-cell derived colonies, with a cosine similarity within the expected 95%
1569 confidence interval (**Methods**), although the rate of C>A mutations appeared to be slightly
1570 elevated (**Extended Data Fig 3c**).

1571

1572 Altogether, sonication followed by Mung Bean nuclease digestion offers an alternative version
1573 of NanoSeq, with considerably lower library yields but error rates in the nano scale ($< 10^{-8}$
1574 errors/bp). This protocol opens the door to applications requiring whole-genome coverage and
1575 to targeted NanoSeq with reliable single-molecule mutation detection.

1576

1577 **Supplementary Note 4 - Modified A-tailing**

1578

1579 After DNA fragmentation, A-tailing of blunt-ended DNA fragments is commonly used in
1580 library preparation protocols before ligation of sequencing adapters. This step involves a DNA
1581 polymerase and dATP, among other reagents. In preliminary experiments (partially digesting
1582 DNA with both HpyCH4V and AluI) we noticed increased levels of C>A and T>A at

1583 restriction sites, and the profile of single-strand consensus, typically caused by DNA damage,
1584 showed an increased amount of G>A, C>A and T>A (**Extended Data Fig 1e**). To explain this
1585 pattern we hypothesised a multi-step mechanism involving: nicking of the DNA duplex by
1586 restriction enzymes (an intermediate step in double-strand cleavage); 3' to 5' exonuclease or
1587 pyrophosphorolysis, during A-tailing, of the dNTP 3' of the nick; incorporation of dATP
1588 opposite C, G or A during A-tailing (causing G>A, C>A or T>A, respectively); and subsequent
1589 sealing of the internal nick during adapter ligation. To block molecules with internal nicks or
1590 gaps, we replaced dATP with a mixture of dATP and ddBTP (ddCTP, ddGTP, ddTTP) during
1591 A-tailing. The presence of internal nicks would trigger DNA polymerase extension until the
1592 incorporation of a ddBTP, making the affected DNA strand unamplifiable. Our results show
1593 that the incorporation of ddBTPs successfully removed artefacts caused by A-tailing
1594 (**Extended Data Fig 1e**).

1595

1596 **Supplementary Note 5 - Chimeric read bundles**

1597

1598 A potential problem in duplex sequencing approaches is the formation of chimeric read bundles
1599 (PCR duplicate families), in which a read bundled contains copies of more than one original
1600 molecule of DNA. This can occur when two original fragments of DNA have identical
1601 breakpoints and barcodes. In such cases a somatic mutation could be undetected because there
1602 is not a consensus at that position in the read bundle, which could result in an underestimation
1603 of the mutation burden. The use of three bp random barcodes in the adaptors at both fragment
1604 ends allow for 4,096 different combinations. With this variability, chimeric read bundles are
1605 expected to be rare with the shallow duplex coverages used in this study.

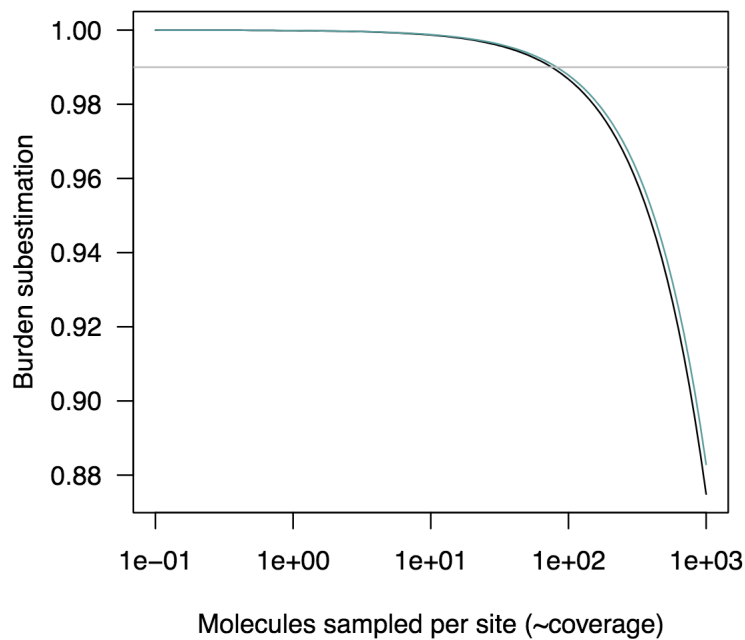
1606

1607 We can study this analytically and empirically for the HpyCH4V protocol. Let c_i be the number
1608 of DNA fragments sampled at a restriction site (i). Since we aim for ~ 3 Gb (1x) of duplex
1609 coverage and we cover $\sim 30\%$ of the human genome, the average c_i is around 3-4 molecules per
1610 restriction site. Let p_j be the vector of relative frequencies of each of the 4,096 barcodes in a
1611 library (we could assume a vector with uniform frequencies $p_j = 1/4,096$ or use empirical
1612 barcode frequencies from a library, which vary modestly). At a given site (i), the probabilities
1613 that one fragment or more than one fragment are tagged with a given barcode (j) can be
1614 modelled as Poisson distributed: $P(x=1, \lambda=c_i p_j)$ and $P(x>1, \lambda=c_i p_j)$, respectively. Assuming
1615 uniform barcode frequencies, the expected fraction of non-chimeric read bundles at a site can
1616 be calculated as: $f_j = P(x=1, \lambda=c_i p_j) / P(x \geq 1, \lambda=c_i p_j)$. Assuming variable barcode frequencies, the
1617 fraction of non-chimeric read bundles expected at a site is a weighted average of this ratio
1618 across barcodes, with the weight of each barcode being proportional to its contribution to
1619 coverage: $w = P(x \geq 1, \lambda=c_i p_j)$. If we conservatively assume that somatic mutations cannot be
1620 called from chimeric reads, f estimates the extent by which the mutation burden (m) may be
1621 underestimated due to chimeric read bundles: $m_{observed} \sim m_{true} f$. Using these equations,
1622 **Supplementary Figure 2** shows that, as expected, chimeric bundles are very rare at the
1623 coverages used in this study, either using uniform or empirically observed barcode frequencies.
1624 In fact, chimeric bundles are expected to be $<5\%$ with whole-genome duplex coverages $<100x$.

1625

1626 To test for the presence of chimeric bundles empirically, we can study the fraction of read
1627 bundles that contain both alleles of a heterozygous SNP. We focused on donor PD43976
1628 (**Supplementary table 1**), for which multiple colonies are available²¹. We ran GATK's
1629 HaplotypeCaller⁵⁵ on each of the colonies and detected 1.4 million reliable heterozygote SNPs
1630 in the donor, defined as those called in at least 90% of the samples and showing a VAF between
1631 0.4 and 0.6. We estimated how many times these heterozygote SNPs passed mapping quality
1632 filters and were seen in two NanoSeq libraries from this donor, and how many times a

1633 consensus call was achieved (requiring at least 90% of the reads from each strand to support
1634 the call). We found that for the two libraries 98.2% and 99.3% of the times when a heterozygote
1635 SNP position was seen, a consensus call could be obtained. This result indicates that
1636 chimerism, if present, must be low. To control for background rates of calling, we calculated
1637 the same numbers for sites surrounding the heterozygote SNP position (-2, -1, +1, +2). For
1638 surrounding sites the proportions were similar, 98.9% and 99.4%. The ratios between
1639 heterozygote and surrounding sites call rates were 0.993 (Poisson CI95% 0.991-0.995, $P =$
1640 $3.8e-14$) and 0.999 (Poisson CI95% 0.997-1.001, $P = 0.46$) for these libraries. Overall, and in
1641 line with theoretical expectation, this analysis indicates that the frequency of chimeric read
1642 bundles and the resulting underestimation of mutation burden is <1% in these libraries.
1643



1644
1645 **Supplementary Figure 2.** Subestimation of the mutation burden due to chimeric read bundles
1646 as a function of coverage per restriction site. This figure shows the subestimation factor (f)
1647 described in the Supplementary Note 5, as a function of coverage per restriction site. The green
1648 line represents f assuming equal frequency of all barcodes ($=1/4,096$) and the black line
1649 represents f using the observed barcode frequency from representative libraries.
1650
1651

1652 **Supplementary Note 6 - Human DNA contamination**

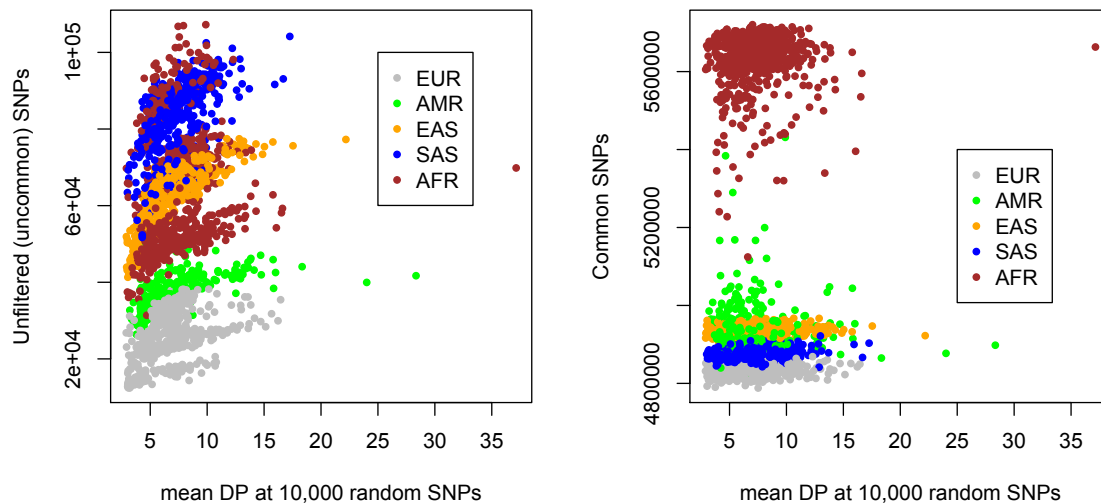
1653
1654 To reduce the impact of contamination on duplex sequencing libraries we generated an
1655 extensive SNP mask (**Methods**). For each NanoSeq library we also estimate the extent of inter-
1656 individual contamination using VerifyBamID²⁴⁶, which we validated using simulations of
1657 contamination fractions as low as 0.1% (**Extended Data Fig 4e,f**).
1658

1659 Applying the SNP mask ($n=26,111,286$) to 1000 Genomes Project data, we estimate that the
1660 mask leaves between 25,666 and 82,765 SNPs unfiltered across samples, with systematic
1661 differences across human populations. Since the 1000 Genomes Project mostly used low-
1662 coverage sequencing and it combined information across samples to help call SNPs in each
1663 sample, the sensitivity to rare SNPs was lower than to common SNPs. This suggests that our
1664 estimates of unfiltered SNPs per sample after applying the SNP mask could be underestimates.

1665 To assess this possibility, we represented the number of unfiltered SNPs in 1000 Genomes
1666 Project samples as a function of their genome coverage and population assignment
1667 (**Supplementary figure 3**). Indeed, and contrary to common SNPs, we found a relationship
1668 between mean depth of coverage and the number of unfiltered alternative alleles
1669 (**Supplementary figure 3**). However, the numbers of unfiltered SNPs appear to plateau above
1670 10x, suggesting that the estimates above are of the right magnitude.

1671
1672 Based on these analyses, we estimate that the rate of SNPs masquerading as somatic mutations
1673 in a NanoSeq sample with 1% of contamination of European ancestry, shifts from 1.6×10^{-6} to
1674 4.0×10^{-8} , after applying the SNP mask. If the contaminant source was the individual with the
1675 highest number of unfiltered SNPs ($n=107,230$, from Luhya ancestry) the false positive rate
1676 would be 1.7×10^{-7} .

1677



1678
1679 **Supplementary Figure 3.** Panels **a** shows the mean coverage depth at 10,000 random SNPs
1680 picked from the set of SNPs not in the common SNP mask. Panel **b** shows the same but for
1681 common SNPs. DP stands for depth of sequencing coverage. EUR stands for European, AMR
1682 for Ad-mixed American, EAS for East Asian, SAS for South Asian, and AFR for African super
1683 populations.
1684

1685
1686 These analyses confirmed that, depending on the mutation burden of the sample and the
1687 ancestry of the contaminant, 1% of contamination can still be problematic after application of
1688 the SNP mask. VerifyBamID is a tool routinely used to estimate human contamination from
1689 sequencing data. The most recent version⁴⁶ is ancestry aware and has been tested for
1690 contamination levels above 1%. Here we performed simulations to evaluate VerifyBamID
1691 performance at 0.1, 0.25, 0.5, 0.75, 1, 2, and 3% contamination levels. To obtain more stable
1692 estimates of contamination we increased the number of markers from 100K to 500K, by
1693 randomly choosing additional SNPs with $MAF > 0.05$ from the 1000 Genomes Project
1694 20130502 release.
1695

1696
1697 We performed two types of simulations; one aimed at evaluating the impact of ancestry and
1698 the other aimed at testing VerifyBamID on NanoSeq data. To evaluate the impact of ancestry,

1699 we mixed BAMs from two individuals from the 1000 Genomes Project using the contamination
1700 fractions specified above. We randomly selected one British individual as the intended sample
1701 (HG00143 GBR/EUR) and 5 other individuals as contaminants: one British (HG00140), and
1702 one from each Africa (NA18867 - YRI), America (HG01060 - PUR), Southern Asia (HG03999
1703 - STU) and Eastern Asia (NA18582 - CHB) continental groups. Our results show that, despite
1704 some deviations, contamination estimates are reasonably accurate for contamination values >
1705 0.1% irrespective of ancestry (**Extended Data Fig 4e**).

1706
1707 Next, we explored how well VerifyBamID works with NanoSeq data. For this experiment we
1708 chose two NanoSeq libraries, one smooth muscle sample from donor PD40794 and one
1709 granulocyte sample from donor PD43980. We simulated contamination of each of the samples
1710 with the other using the contamination levels 0.1, 0.25, 0.5, 0.75, 1, 2, and 3%, as above.
1711 Results support the ability of VerifyBamID to detect levels of contamination > 0.1% for
1712 NanoSeq data (**Extended Data Fig 4f**).

1713 1714 **Supplementary Note 7 - Further details on the estimation of mutation burden in standard** 1715 **sequencing data**

1716
1717 Mutation burden estimation in NanoSeq is unaffected by the clonality of a sample or the depth
1718 of coverage. In contrast, the somatic mutation calling sensitivity in standard sequencing data
1719 depends on both clonality and coverage. Except for standard sequencing of smooth muscle (for
1720 which we did not attempt to correct the mutation burden), all of the samples compared here
1721 were clonal or nearly clonal, but their sequencing coverage was still variable. Somatic
1722 mutations occurring at genomic regions with low coverage are more likely to be undetected.
1723 To estimate the sensitivity of CaVEMan we simulated 10,000 clonal heterozygous mutations
1724 (VAF~50%) in seven BAM files using *bamsurgeon*⁵⁶ (with parameters "--ignore-snps --insane
1725 --aligner mem"). Of the 10,000 mutations requested, *bamsurgeon* successfully simulated
1726 around 9,000 in each sample. For those mutations successfully simulated we found that in
1727 regions with at least 20x coverage, very few mutations were missed by CaVEMan (99.83%
1728 sensitivity across the seven BAM files). After application of the various filters used on
1729 CaVEMan calls, sensitivity dropped to 96.42%. The filters removing most simulated mutations
1730 were the panel of normals (VUM; 2.2%), the simple repeats filter (SR; 1.0%) and the
1731 centromeric repeats filters (CR; 0.2%). Based on this high sensitivity, we decided to restrict all
1732 comparisons of mutation burden between standard sequencing and NanoSeq in the study to the
1733 fraction of the genome covered by at least 20 reads in each sample.

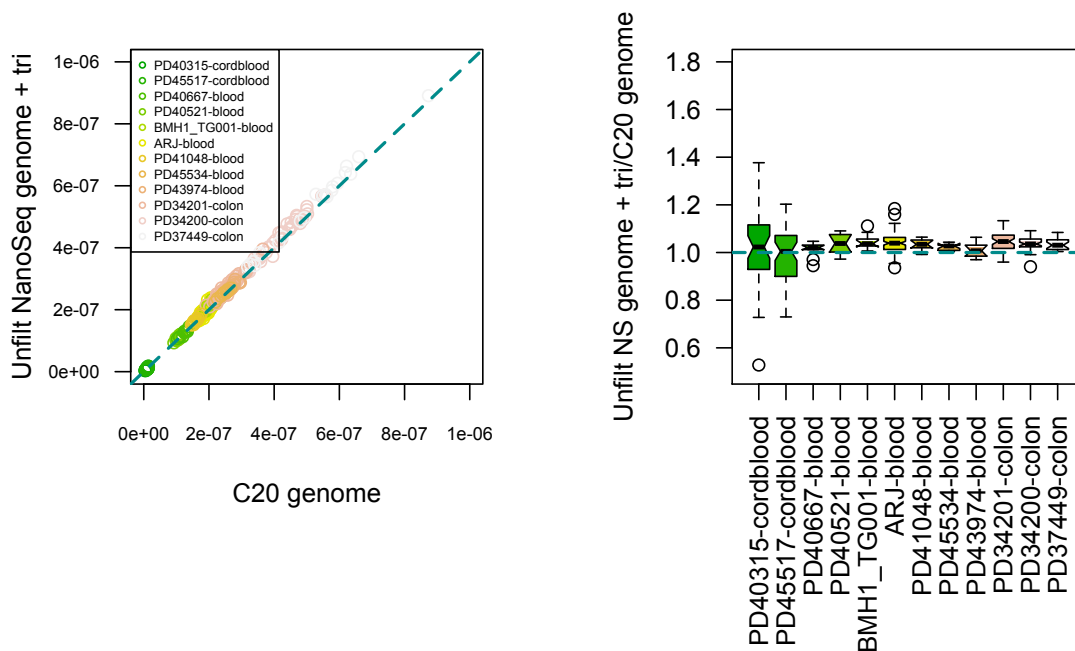
1734
1735 Another important consideration to compare the two protocols comes from the fact that the
1736 NanoSeq coverage is uneven across the genome due both to the use of restriction enzymes and
1737 the application of stringent filters. With our choices of restriction enzyme (HpyCH4V) and size
1738 selection (250-500 bps), about 27% of the genome is covered. Since mutation rates are known
1739 to vary across the genome, to avoid systematic biases we decided to further restrict the
1740 comparison of standard sequencing data and NanoSeq to regions covered by NanoSeq and
1741 considered callable. The *NanoSeq genome* (*g*) was defined using a sample with high NanoSeq
1742 coverage (PD43976 / 33796#41; **Supplementary Table 1**), and including only sites covered
1743 by at least one read bundle and passing all our filters (*g* = 783,199,533 bp).

1744
1745 In summary, for the final comparison between CaVEMan and NanoSeq, we focused on the
1746 fraction of the genome (*T*) overlapping the *NanoSeq genome* (*g*) and having at least 20x
1747 coverage (*c*) in each sample, i.e. $T = g \cap c$. Low-coverage samples with $T < 200$ Mb were not
1748 analysed further.

1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770

Mutation calls falling in the comparable genome fraction (T) were identified (m) and a mutation rate (r) was calculated as $r = m / T$, with associated 95% Poisson confidence intervals. Given the differences in trinucleotide sequence composition between the whole reference genome and the NanoSeq genome, we corrected the observed mutation rates as described in Methods (**Correction of mutation burden and trinucleotide substitution profiles**), resulting in r' . Corrected confidence intervals were calculated as $CI' = CI * r' / r$. To estimate the total number of mutations per cell (M), we multiplied r' (and its associated confidence intervals CI') by the size of the callable diploid genome (D_g), taken here as 5,722,652,910 base pairs (and half of this for the haploid genome of sperm cells).

We found that the corrected mutation rates (r') on T were consistently ~20% higher than estimates based on c (the fraction of the genome with at least 20x coverage; **Extended Data Fig 4a,b**), the latter defined as $r_c = m_c * D_g / c$, where m_c is the number of mutations in c . To determine whether the 20% increase is due to uneven NanoSeq coverage or to NanoSeq stringent filters, we estimated the corrected rates in T' , defined as the fraction of the genome covered by NanoSeq but without applying our strict mapping quality filters. The results show that, while the rates in T are 20% higher than in c , the rates in T' do not increase considerably (**Supplementary Figure 4**). This indicates that the higher rates obtained with NanoSeq are caused by limited calling sensitivity in standard sequencing data in the regions filtered out by NanoSeq. That is, traditional mutation burden estimates with standard sequencing technologies are likely underestimates due to low sensitivity in certain genomic regions.



1771
 1772
 1773
 1774
 1775
 1776

Supplementary Figure 4. Mutation rates in the unfiltered NanoSeq genome (with coverage ≥ 20) compared to mutation rates in the fraction of the genome with at least 20x coverage (left), and the ratio of the two (right). By comparing this to **Extended Data Fig 5a,b** it becomes clear that the increase observed in the NanoSeq genome is mainly due to our mapping quality filters.

1777

1778 **Supplementary Note 8 - Validation of indel calls**

1779

1780 To estimate the indel error rate we compared Pindel calls⁵⁰ in single-cell derived cord blood
1781 colonies to our indel calls in NanoSeq cord blood granulocytes. Comparing indel rates is
1782 particularly difficult given the known problems of specificity and sensitivity associated with
1783 indel calling. For this comparison we applied the same approach that we used to compare
1784 CaVEMan and NanoSeq, restricting the analysis to regions with at least 20x coverage and
1785 falling in the NanoSeq-covered genome. Pindel estimated 6.5×10^{-10} indels / bp (CI95% $4.8 \times$
1786 10^{-10} - 8.7×10^{-10}), while NanoSeq estimated 1.8×10^{-9} indels / bp (CI95% 1.2×10^{-9} - $2.6 \times$
1787 10^{-9} ; **Extended Data Fig 5c**). Although NanoSeq estimates are higher, and some of this
1788 difference may be due to higher rates of indels in differentiated cells, we can confidently
1789 estimate an indel error rate for NanoSeq $<3 \times 10^{-9}$ / bp.

1790

1791 The reliability of our indel calls is further supported by the linear accumulation of indels with
1792 age observed for granulocytes, smooth muscle and neurons. To further investigate the quality
1793 of our indel calls we also compared indel profiles for samples with reliable indel calls from
1794 standard whole-genome sequencing data. These included a bladder tumour, colonic crypts
1795 (with and without the colibactin signature), and *POLE* and *POLD1* mutants. NanoSeq indel
1796 profiles matched closely the reported indel profiles for colibactin²⁷, *POLE*/*POLD1* samples⁴⁸,
1797 and a previously published bladder tumour sample³⁴ (**Extended Data Fig 5d**).

1798

1799 Indel profiles for each cell type analysed in this manuscript are shown in **Fig 3d,m** and
1800 **Extended Data Fig 9c**.

1801

1802 **Supplementary Note 9. Estimation of the minimum number of divisions required to** 1803 **produce granulocytes from HSCs.**

1804

1805 Estimates of the population size of human haematopoietic stem cells (HSC) range from 25,000
1806 to 1.3 million and the human HSC division rate is estimated to be between one per 28 days and
1807 one per 4 years^{21,57-59}. The relatively small population size and division rates of HSC contrast
1808 with the staggering production of blood cells throughout life. On the order of 10^{11} granulocytes
1809 are estimated to be produced every day⁶⁰ and on the order of 1.4×10^{14} blood cells are estimated
1810 to be produced every year considering all mature cell types⁵⁸.

1811

1812 In mouse and cat, the enormous net amplification during blood cell production is achieved by
1813 means of between 17 and 19.5 effective cell divisions^{61,62}. Accurate estimates in humans are
1814 difficult to obtain, however, we can calculate a lower bound for the number of cell divisions
1815 required based on the size of the stem cell population and the number of differentiated cells
1816 produced.

1817

1818 Theoretically, the minimum number of divisions required to produce N differentiated cells
1819 from a single cell is achieved by a perfectly bifurcating tree⁶³, in which a single cell expands
1820 into $N = 2^d$ cells. Hence, the minimum number of cell divisions that must separate a HSC from
1821 an average differentiated cell can be calculated as $d = \log_2 N$. If we assume, as an example,
1822 that the HSC pool in humans is 100,000 cells and that 10^{11} granulocytes are produced every
1823 day, then the minimum number of cell divisions with a perfect bifurcating tree would be $d =$
1824 $\log_2(10^{11}/10^5)$, that is, at least 19.9 cell divisions. However, we know that HSC divide
1825 infrequently (around once per year) and need to self-sustain. To maintain homeostasis, on
1826 average, the division of a HSC results in a HSC and a progenitor. Because progenitors have to

1827 produce 10^{11} granulocytes every day for a year (assuming an average division rate of one
1828 division per year), we have to consider the total number of granulocyte production during that
1829 period, making $d = \log_2(3.65 \cdot 10^{13}/10^5)$, i.e. $d \geq 28.4$ divisions assuming a HSC division
1830 rate of 1/40 weeks. This is a theoretical lower bound estimate of d , because it assumes an
1831 optimum bifurcating lineage and because it does not consider the production of other blood
1832 cell types. Although estimates of the number of HSCs and the HSC division rate in humans
1833 vary considerably, even the most extreme estimates (1.3 million HSCs dividing every 28 days)
1834 predict $d \geq 20$.

1835
1836 Our linear regression model estimated a difference between the intercepts of granulocytes and
1837 HSC/MPPs of ~ 58 mutations, although the difference was not significantly different from 0
1838 (CI95%: -13.1-121.1, **Fig 2b**). Based on this estimation of the difference of mutations between
1839 granulocytes and HSC/MPPs, we can estimate an upper bound of ~ 2 mutations (58/28) per cell
1840 division during transient proliferation and differentiation. Given that HSCs accumulate 19.8
1841 mutations per year and divide on the order of once per year, these estimates suggest that only
1842 a small minority of mutations in HSCs are likely to represent replication errors. Alternatively,
1843 to explain the small difference observed in mutation burden between HSC/MPPs and
1844 granulocytes as a function of replication-associated mutagenesis alone, and taking $d \geq 28.4$,
1845 HSC/MPPs would need to divide >10 times per year or have a mutation rate per division >10
1846 times higher than that of transient progenitors.

1847
1848 Altogether, if we assume that HSC/MPPs divide infrequently and are at least as protected from
1849 mutagenesis as transient progenitor cells, the observed mutation burden data suggests that most
1850 mutations in HSC/MPPs accumulate non-replicatively, as a function of time rather than cell
1851 division.

1852

1853

1854 **Supplementary Note 9 - Comparison between smooth muscle and single-cell derived** 1855 **colonies of skeletal muscle satellite cells**

1856

1857 To our knowledge, our manuscript contains the first description of the mutational landscape
1858 of smooth muscle. A previous study described the mutation rates and mutational spectrum of
1859 skeletal muscle satellite cells by expanding single satellite cells into colonies in vitro¹¹. In
1860 that study, the authors sequenced colonies from young and old donors, as well as colonies
1861 grown in vitro for different lengths of time, to quantify the effects of in vitro culture.

1862

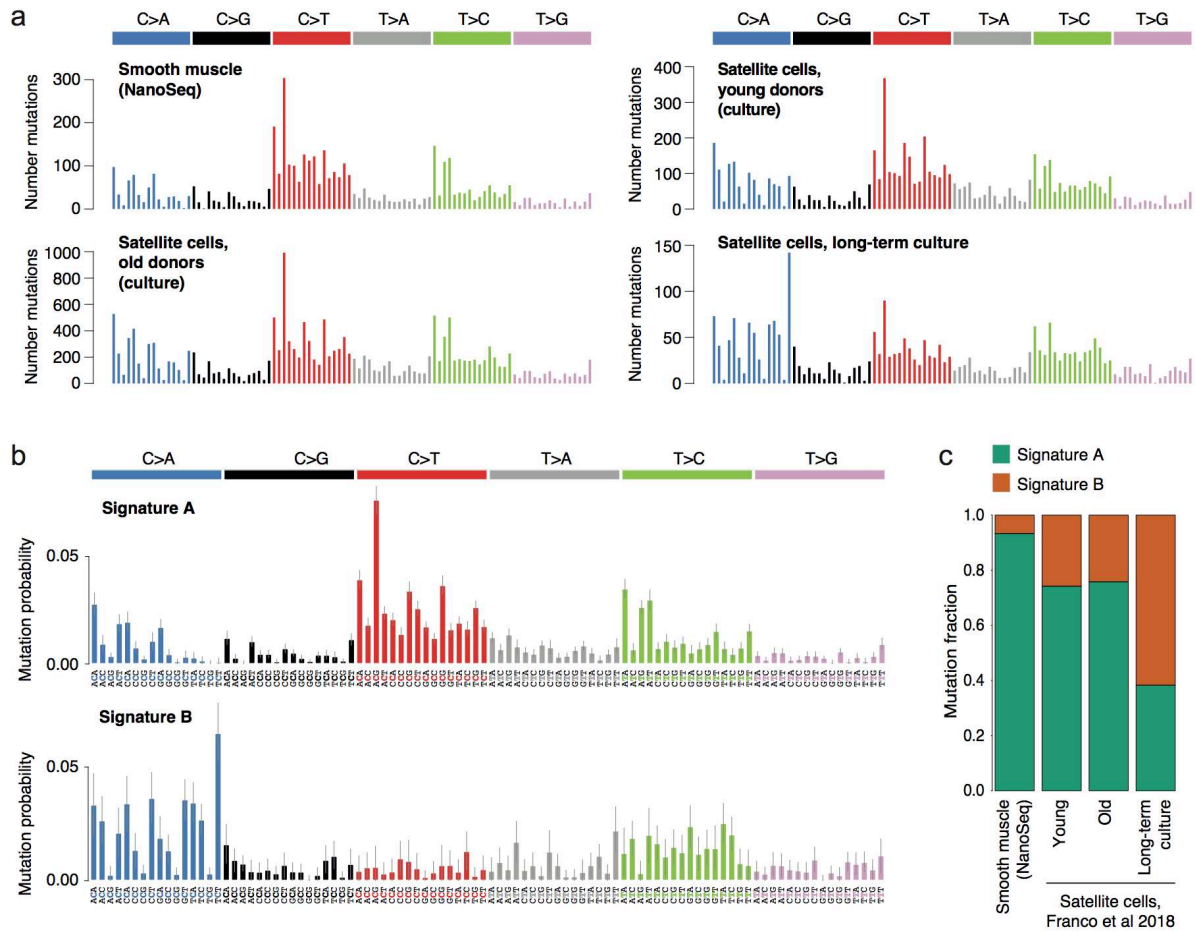
1863 Owing to differences in mutation calling sensitivity, comparison of mutation burdens
1864 estimated with different sequencing strategies and coverages is challenging from public
1865 mutation data, but mutational spectra can be more easily compared. We find that the
1866 substitution profile of smooth muscle is remarkably similar to that of satellite cells (cosine
1867 similarities of 0.96 for young and old donors; **Supplementary figure 5a**). The similarity is,
1868 however, markedly lower with the long-term cultured colonies (cosine of 0.80);

1869 **Supplementary figure 5a**).

1870

1871 Using sigfit we extracted two signatures from the set of four substitution profiles composed
1872 of NanoSeq smooth muscle, satellite cells from young and old donors, and long-term cultured
1873 colonies (**Supplementary figure 5b**). Signature A is very similar to NanoSeq smooth muscle
1874 (cosine of 0.99), whereas signature B is similar (cosine of 0.84) to signature C in Blokzijl et
1875 al¹⁰, which is associated to mutations introduced during in vitro culture. This signature has a

1876 greater contribution in satellite cells compared to smooth muscle, and increases in long-term
 1877 cultured colonies (**Supplementary figure 5c**).
 1878
 1879



1880 **Supplementary figure 5. a**, Substitution profiles of smooth muscle (top left) and satellite
 1881 cells from young and old donors, and in long-term culture¹¹. **b**, Two extracted signatures
 1882 from satellite cells. **c**, Estimated exposure of groups of samples to each of the two extracted
 1883 signatures, showing how signature B contribution becomes stronger in long-term culture and
 1884 is practically absent from NanoSeq data.
 1885