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

Single-molecule mutation detection unravels the mutational landscapes of differentiated cells

Federico Abascal[1], Luke M. R. Harvey[1,#], Emily Mitchell[1,2,#], Andrew R. J. Lawson[1,#], Stefanie V. Lensing[1,#], Peter Ellis[1,3,#], Andrew J. C. Russell[1], Raul E. Alcantara[1], Adrian Baez-Ortega[1], Yichen Wang[1], Eugene Jing Kwa[1], Henry Lee-Six[1], Alex Cagan[1], Tim H. H. Coorens[1], Michael Spencer Chapman[1], Sigurgeir Olafsson[1], Steven Leonard[1], David Jones[1], Heather E. Machado[1], Megan Davies[2], Nina F. Øbro[2,4], Krishnaa Mahubani[5,6], Kieren Allinson[7], Moritz Gerstung[8], Kourosh Saeb-Parsy[5,6], David G. Kent[2,9], Elisa Laurenti[2,4], Michael R. Stratton[1], Raheleh Rahbari[1], Peter J. Campbell[1,4], Robert J. Osborne[1,10,*], Iñigo Martincorena[1,*].

# These authors contributed equally
* Corresponding authors: *r.osborne@biofidelity.com* (R.J.O.), *im3@sanger.ac.uk* (I.M.)

Affiliations:
[1] Wellcome Sanger Institute, Hinxton CB10 1SA, UK.
[2] Wellcome - MRC Cambridge Stem Cell Institute, Cambridge Biomedical Campus, Cambridge CB2 0AW, UK.
[3] Current address: Inivata, Glenn Berge Building, Babraham Research Campus, Babraham, Cambridge, CB22 3FH, UK
[4] Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK.
[5] Department of Surgery, University of Cambridge, Cambridge CB2 0QQ, UK.
[6] NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK.
[7] Cambridge Brain Bank, Division of the Human Research Tissue Bank, Box 235, Level 5, Addenbrooke's Hospital, Hills Rd, Cambridge, CB2 0QQ, UK.
[8] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK.
[9] York Biomedical Research Institute, Department of Biology, University of York, York YO10 5DD, UK.
[10] Current address: Biofidelity, 330 Cambridge Science Park, Milton Road, Cambridge, CB4 0WN, UK

Abstract: Somatic mutations drive cancer development and may contribute to ageing and other diseases[1,2]. Yet, the difficulty of detecting mutations present only in single cells or small clones has limited our knowledge of somatic mutagenesis to a minority of tissues. To overcome these limitations, we introduce nanorate sequencing (NanoSeq), a new duplex sequencing protocol that avoids end-repair-associated errors to achieve mutation detection error rates <5 errors per billion base pairs in single DNA molecules from populations of cells. This rate is two orders of magnitude lower than typical somatic mutation loads, enabling the study of somatic mutation in any tissue independently of clonality. We exploit the single-molecule sensitivity of NanoSeq to study somatic mutations in non-dividing cells across several tissues, comparing stem cells to differentiated cells and studying mutagenesis in the absence of cell division. Differentiated cells in blood and colon displayed remarkably similar mutation loads and signatures to their corresponding stem cells, despite mature blood cells having undergone a considerable number of additional cell divisions. We then characterised the mutational landscape of post-mitotic neurons and polyclonal smooth muscle. This confirmed that neurons accumulate somatic mutations at a constant rate throughout life in the absence of cell division, with similar mutation

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

Introduction

Somatic mutations occur in our cells as we age, driving cancer development and potentially contributing to ageing and other diseases. Despite their importance, their study remains challenging due to technical limitations. Because any given somatic mutation in a normal tissue is typically present in a small group of cells or even in a single cell, detecting them requires special approaches, such as ultra-deep sequencing of small biopsies[3-5], laser microdissection[6-8], isolation of single-cells followed by in vitro expansion into organoids or colonies[9-11], or single-cell sequencing[12-14]. While these technologies are changing our understanding of somatic mutagenesis, the error rate of single-cell sequencing remains high[15], and other approaches are typically limited to mitotically-active cell types.

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

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

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

Duplex sequencing has a theoretical error rate $<10^{-9}$ errors/bp, which is the probability of two early and complementary PCR errors in both strands[16]. Given that this theoretical limit is lower than the typical mutational load of human tissues, it raises the possibility of quantifying somatic mutation rates in any cell type, independently of its clonal architecture. This is the rationale of BotSeqS, a whole-genome duplex sequencing protocol[19]. In practice, however, mapping errors and the accidental copying of errors between strands during library preparation violate the independence of both strands and limit the accuracy of duplex sequencing[19,20]. The actual error rates of duplex sequencing protocols have remained difficult to measure[16], but some protocols

report error rates above 10^{-7} errors/bp[20], translating into hundreds to thousands of errors per diploid genome.

A difficulty in measuring the error rate of existing duplex sequencing protocols has been the lack of control samples with known mutation rates. To evaluate the performance of the existing BotSeqS protocol (**Fig 1a**), we first analysed a sample of granulocytes from a 59-year-old donor from whom 110 single-cell derived blood colonies had been whole-genome sequenced[21] (**Supplementary Table 1,2**). We found that the estimates of mutation burden per diploid genome from BotSeqS were two-fold higher than those from the colonies (**Fig 1b**), and that the substitution profiles were dissimilar (cosine similarity of 0.71; **Fig 1c**), with increased C>A and C>G substitution rates. Analysing the distribution of substitutions across the reads revealed a large excess of G>T/C substitutions near the 5'ends of DNA fragments, and an imbalance over C>A/G substitutions that affected the entire read length (**Fig. 1d** and **Extended Data Figures 1 and 2**). These substitution imbalances are incompatible with real mutations and reflect errors introduced during library preparation[22] (**Methods**, **Supplementary Note 1**). We confirmed that the same imbalances, together with an additional C>T asymmetry, were present in the original BotSeqS publication[19] (**Fig 1d**). Extensive trimming of read ends only partially alleviated these errors (Extended Data Fig 2). Based on these results, we estimate that BotSeqS introduced ~1,500 errors per diploid genome in our samples, equivalent to an error rate ~2.6 x 10^{-7} errors/bp.

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

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

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

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

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

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

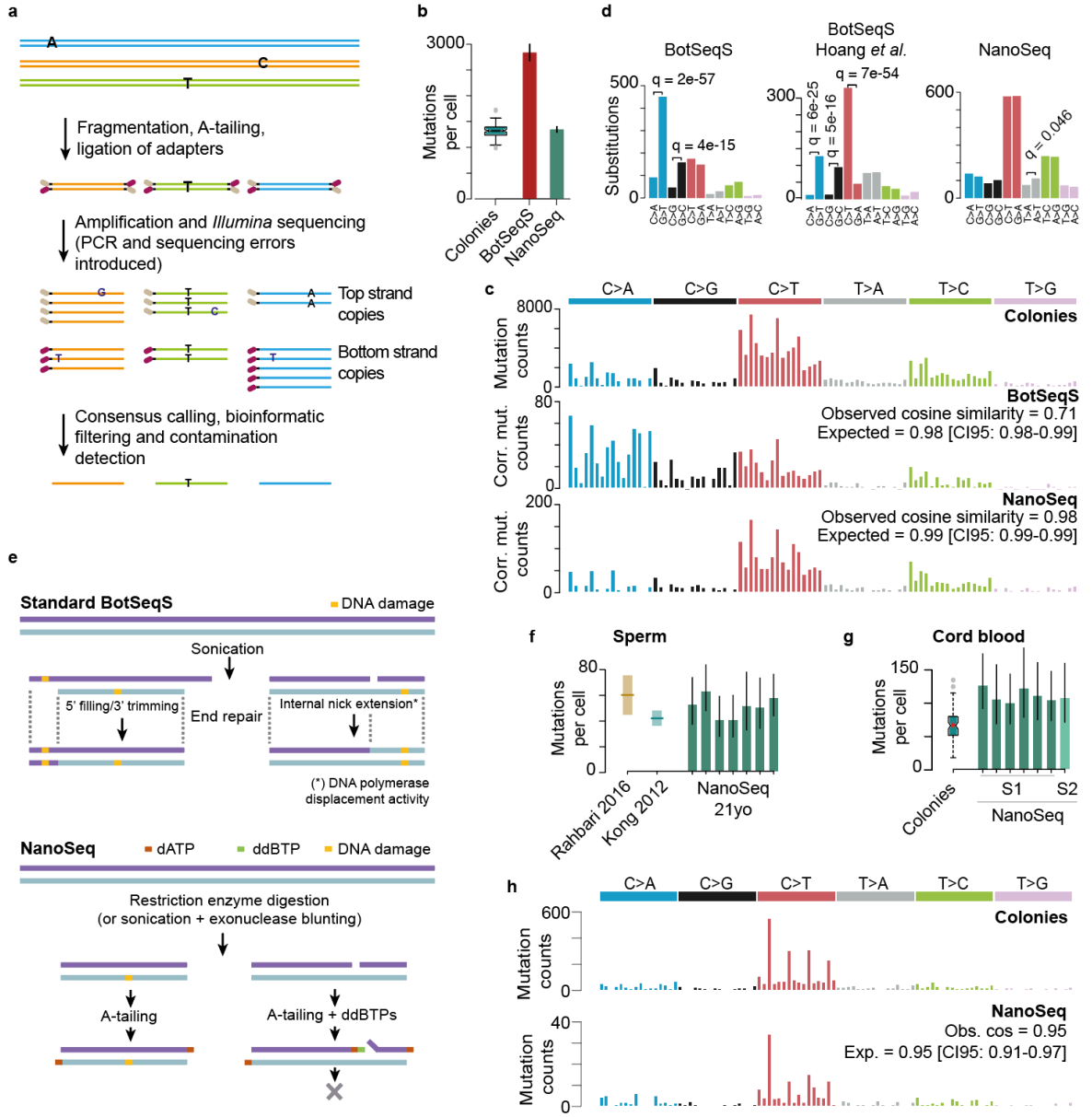


Figure 1 | Standard BotSeqS and NanoSeq sequencing protocols. **a**, Fundamentals of duplex sequencing protocols. **b**, Mutation burden estimates in granulocytes using BotSeqS and NanoSeq, compared to standard results with single-cell derived blood colonies. Box plot show the interquartile range, median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

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 cells[25], differentiated cells could conceivably have higher mutational loads and undescribed mutational signatures[14].

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 donor[21] (**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 ~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

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

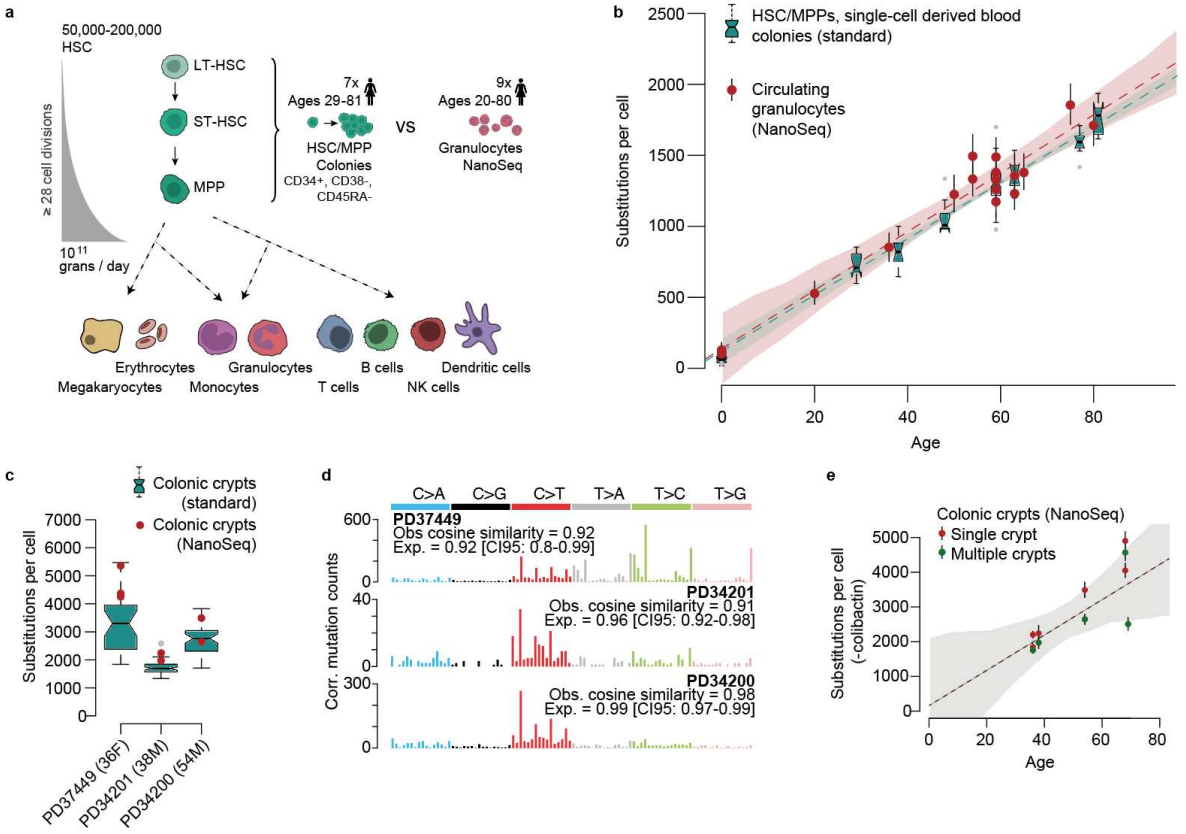


Figure 2 | Mutation analyses of differentiated cells. **a**, Schematic representation of the hematopoietic lineage showing which cell types and donors were analysed. **b**, Substitutions per cell for donors of different ages, comparing estimates from NanoSeq granulocytes (red) to standard sequencing of single-cell derived blood colonies (dark cyan boxplots); boxplots and confidence intervals as in **Fig 1b**; red and dark cyan dashed lines are linear mixed regression models; linear mixed model 95% confidence intervals for NanoSeq data calculated through parametric bootstrapping. For granulocytes, the intercept is 137.6 [CI95% -117.6-413.2] and the slope 20.6 [15.8-25.2]. For blood colonies, the intercept is 120.4 [27.9-218.5] and the slope 19.8 [18.2-21.4]. **c**, Comparison between standard methods and NanoSeq burden estimates for colonic crypts from three donors. **d**, Substitution profiles for colonic crypts from the three donors and cosine similarities to profiles obtained with standard methods. **e**, Accumulation of substitutions throughout life in colonic crypts from 5 donors, excluding substitutions attributed to the episodic colibactin signature; confidence intervals 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 intercept).

To extend the comparison of stem cells and differentiated cells to another tissue with a well-understood stem cell organisation, we studied colonic epithelium. Estimates of the somatic mutation rate in colonic stem cells are available from whole-genome sequencing of clonal organoids derived from Lgr5+ cells[10] and from sequencing single colonic crypts[6]. Genome sequencing of whole crypts can be used to estimate the somatic mutation rate of colonic stem cells, as colonic crypts are clonally derived from a single stem cell. However, the process of reaching clonality through genetic drift in the population of stem cells within a crypt is estimated to take several years in humans[26], which could lead to an underestimation of mutation burdens using single-crypt sequencing.

For three previously-studied donors we compared standard whole-genome sequencing of laser-microdissected colonic crypts[6] to NanoSeq data from single crypts or groups of crypts. This revealed similar estimates of mutation burden, despite the lag to clonality in standard sequencing (**Fig 2c**). Mutational burden and signatures from differentiated cells in colonic epithelium were overall consistent with those found by previous studies on colonic stem cells, with a dominance of SBS1, SBS5 and, in some donors, a colibactin signature[27] (**Fig 2d,e**).

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

Lifelong mutagenesis in post-mitotic neurons and polyclonal smooth muscle

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

We applied NanoSeq to frontal cortex neurons from 8 healthy donors and 9 Alzheimer's disease (AD) patients (**Supplementary Table 1**), using nuclei sorting with the *NeuN* neuronal marker (**Methods**; **Extended Data Fig 7a**). These data revealed a tight linear accumulation of 20.0 substitutions (linear regression, CI95%:19.1-20.9) and 3.1 indels (CI95%:2.9-3.3) per year, approximately constant throughout life (**Fig 3a,b**). This confirms that mutations accumulate in a clock-like fashion in cortical neurons, in the absence of cell division, consistent with observations from single-cell sequencing[13].

These data shed new light on previously published single-neuron sequencing results. A study using SNP-phased error-corrected single-cell sequencing reported three dominant signatures in neurons, one that increased linearly with age and two that did not[13]. The spectrum found by NanoSeq, the burden per genome and the mutation rate per year closely resemble the age-associated signature in that study (cosine similarity 0.96; **Extended Data Fig 7b,c**). The other two mutational signatures, responsible for around 72% of all mutations reported in the study and highly variable across single-cell libraries (**Extended Data Fig 7d**), appear exclusively in single-cell data and seem more consistent with amplification errors or transient DNA damage. Consistent with this hypothesis, the dominant signature in single-neuron data closely resembles a single-cell-specific signature reported in vitro[15] (cosine similarity 0.97, **Extended Data Fig 7b**).

To better understand the mutational processes active in neurons in the absence of cell division, we carried out signature decomposition on NanoSeq data from neurons together with data from granulocytes, colonic crypts and smooth muscle (described below). Three signatures were extracted (**Fig 3e**): signatures A and C imperfectly resembled SBS5 (cosine similarity 0.80) and SBS16 (0.78), respectively, while signature B closely matched SBS1 (C>T changes at CpG dinucleotides, cosine similarity 0.96). It is conceivable that SBS5, which appears to be a ubiquitous signature in normal tissues and cancer genomes[30], reflects a collection of co-

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

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

Indel analysis revealed a higher relative frequency of indels in neurons than in other tissues, caused by an unusual signature characterised by indels longer than 1bp (**Fig 3d,m**; **Extended Data Fig 9c**). This indel signature and its association with highly expressed genes has some resemblance to a little-understood mutational process recently described in cancer genomes[31] (**Extended Data Fig 9d**).

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

To extend these analyses to another tissue not amenable to standard sequencing methods, we studied smooth muscle. Visceral smooth muscle cells are believed to divide infrequently in normal conditions[32]. Using laser microdissection, we collected samples of smooth muscle from 10 donors and from two different organs, bladder and colon (**Supplementary Table 1,2**; **Extended Data Fig 6b, 10a**). As expected for a polyclonal tissue, standard whole-genome sequencing detected few mutations and at low allele frequencies in these samples (**Extended Data Fig 10b**,**c**, **Methods**). In contrast, NanoSeq revealed that the substitution and indel burdens increase linearly with age, with ~24.7 substitutions per year per diploid genome (CI95%:22.5-27.0) and ~2.1 indels per year (95%:1.7-2.5) (**Fig 3j,k**). Despite their different anatomical origin, smooth muscle cells from the bladder and colon walls showed relatively similar mutation rates (mixed-effects linear regression, P = 0.6 for substitutions, P = 0.04 for indels).

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

Data Fig 7f), with similar contributions in smooth muscle from bladder and colon and across donors (**Fig 3n**). The smooth muscle spectra also resemble that of skeletal muscle satellite cells, studied by in vitro expansion[11] (**Supplementary Note 10)**.

Altogether, granulocytes, smooth muscle and neurons showed more limited variation in mutation rate and spectra across individuals than has been observed in epithelia exposed to exogenous mutagens, such as skin[3], colon[6] (**Fig 2c**), bronchus[33] or bladder[8,34]. This suggests that the rate of endogenous mutagenesis across individuals is modest, at least in the cohort studied. The observation of a linear accumulation of mutations in post-mitotic neurons, with similar burdens and signatures to some mitotically active tissues, suggest that dominant mutational processes observed across tissues may act independently of cell division.

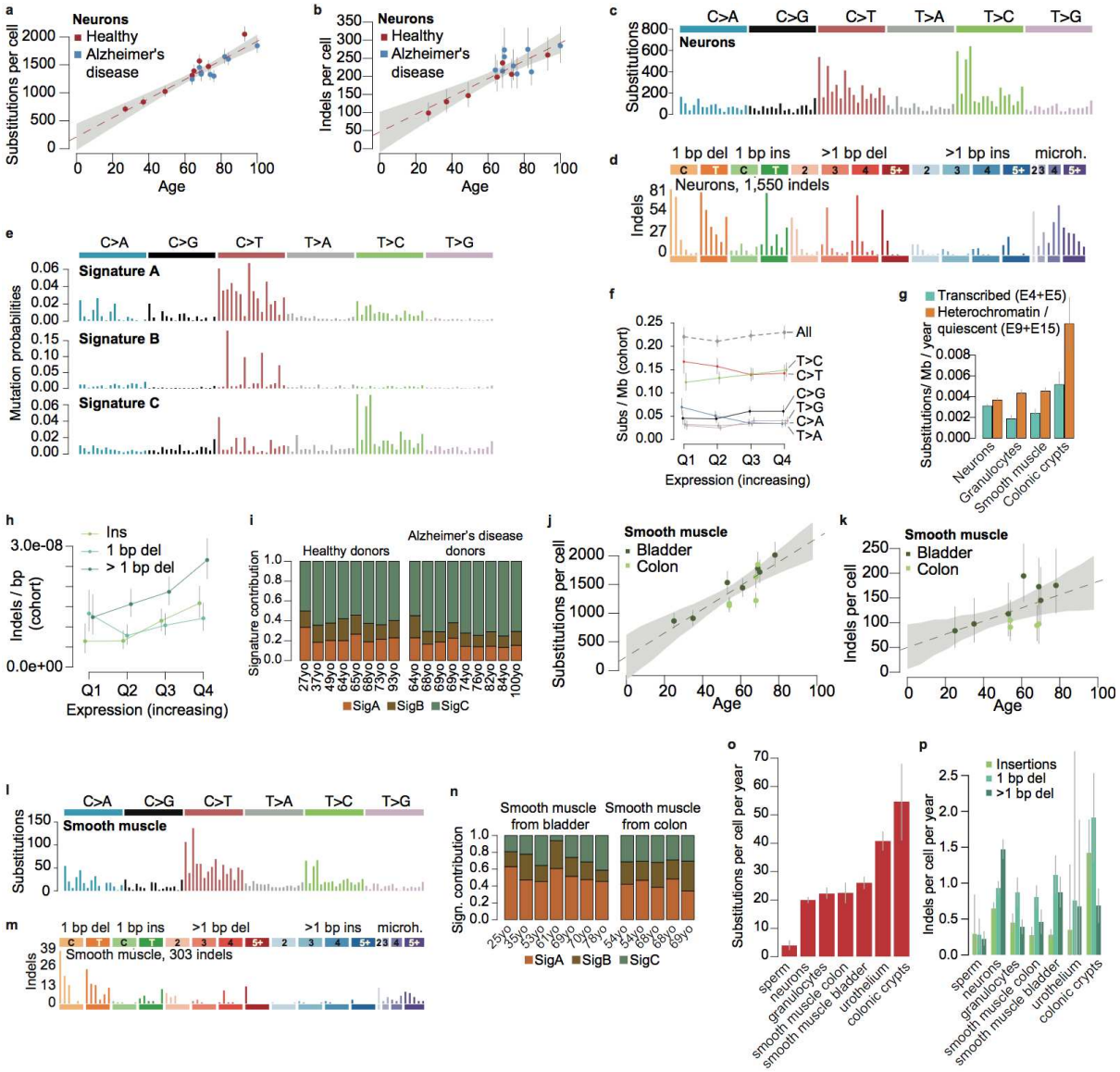


Figure 3 | Mutation landscape in neurons and smooth muscle. **a-b**, Substitution and indel accumulation per neuron throughout life; point estimate confidence intervals as in **Fig 2a**; grey area shows simple linear model 95% confidence intervals. Intercept and slope for substitutions: 210.5 [-26.9-448.0]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

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

Discussion

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

Most of our current knowledge of somatic mutagenesis is restricted to mitotically-active cells. We have exploited the ability to sequence any cell type to explore the mutational landscape of non-dividing cells in a diversity of mitotically-active or inactive tissues. This has enabled us to compare the mutational landscape of differentiated cells and stem cells in blood and colon, and to study somatic mutagenesis in the absence of cell division. A remarkable observation that emerges from these data is that somatic mutation rates vary modestly (~2-3 fold) across a diverse range of somatic cell types, largely independently of cell division rates (**Fig 3o,p**, **Suppl. Note 6**). Indeed, similar mutation rates are found in non-dividing cortical neurons, in smooth muscle and in blood; or in colonic epithelium, which divides every few days, and in mostly quiescent hepatocytes[10] or urothelial cells (**Fig 3o,p**).

DNA replication and cell division have long been assumed to be major sources of somatic mutations, either due to DNA polymerase errors or the fixation of unrepaired damage during replication[35]. However, the linear accumulation of somatic mutations in post-mitotic neurons confirms that dominant mutational processes can occur independently of cell division. These mutations may result from the interplay between endogenous DNA damage and repair that cells are engaged in at all times. The similar mutation burden and signatures in granulocytes and in the stem cells responsible for long-term maintenance of blood, despite a different divisional load, could also be consistent with a time-dependent rather than a division-dependent accumulation of somatic mutations during haematopoiesis. Altogether, it is conceivable that division-independent mutational processes play a larger role in adult somatic mutagenesis than it is commonly assumed.

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Methods

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

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

A sample of granulocytes from a 59-year-old male donor (PD43976_59yo) from whom 110 HSPC colonies were available[21] was used for initial validation of the BotSeqS and NanoSeq protocols (**Supplementary Tables 1,2**). To estimate the NanoSeq error rate, cord blood granulocytes from two neonatal donors were sequenced by NanoSeq and the mutation burdens and spectra compared to those from 50 HSC/MPP colonies per donor. For the comparison of differentiated and stem cells, NanoSeq data from granulocytes from 9 donors of different ages was compared to standard sequencing of single-cell derived HSC/MPP colonies from 6 donors (10 HSC/MPP colonies per donor) and 110 HSPC colonies already available from a 59-year-old donor[21]. These 110 HSPC included 67 HSC/MPPs, 32 megakaryocyte–erythrocyte progenitors (MEP), 7 granulocyte–macrophage progenitors (GMP) and 4 common myeloid progenitors (CMP).

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

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

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

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

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

DNA extraction was performed using the DNeasy 96 blood and tissue plate kit (Qiagen) for larger HSC colonies, or the Arcturus Picopure DNA Extraction kit (ThermoFisher) for smaller HSC colonies. Both kits were used as per the manufacturer's instructions. Extracted DNA (1-5ng) from each colony was processed using a recently developed low-input enzymatic fragmentation-based library preparation method[38]. All samples were subjected to whole genome sequencing at 8-35X coverage on either the HiSeq X or the NovaSeq platforms (Illumina) to generate 150 bp paired-end reads. BWA *mem* was used to align sequences to the human reference genome (NCBI build37).

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

Sperm samples

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

Laser microdissection of colonic crypts and bladder/colon smooth muscle

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

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

Prior to laser-capture microdissection, samples were processed, embedded in paraffin and sectioned as described previously[34]. Microbiopsies were dissected using an LMD7 microscope (Leica Microsystems). Examples of microdissected regions for both specimen types can be found in **Extended Data Figures 6** and **10**. Proteolysis of isolated regions was performed using

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

Neuron nuclei sorting from frontal cortex samples

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

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

BotSeqS and NanoSeq library preparation protocols

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

NanoSeq libraries were prepared as follows: 10 ng of genomic DNA or LCM cut sections in 20 µl buffer were purified using 100 µl of a 50:50 water and AMPure XP bead mixture and eluted in 20 µl nuclease free water. 20 µl of the bead suspension was taken forward into an on-bead fragmentation reaction. Fragmentation occurred in a final volume of 25 µl including 2.5 µl 10x CutSmart buffer (500 mM Potassium Acetate, 200 mM Tris-acetate, 100 mM Magnesium Acetate, 1 mg/ml BSA, pH 7.9 at 25°C), 0.5 µl 5 U/µl HpyCH4V (**Supplementary Note 2**), and 2 µl dH_20. Fragmentation reactions were incubated at 37 ºC for 15 min, purified with 2.5x AMPure XP beads and resuspended in 15 µl nuclease-free water. Fragmented DNA was A-tailed in 15 µl reactions including 10 µl fragmentation product, 1.5 µl 10x NEBuffer 4 (500 mM Potassium Acetate, 200 mM Tris-acetate, 100 mM Magnesium Acetate, 10 mM DTT, pH 7.9 at 25°C), 0.15 µl 5 U/µl Klenow fragment (3′→5′ exo-, New England Biolabs), either 1.5 µl 1 mM dATP or 1.5 µl 1 mM dATP/ddBTPs (**Supplementary Note 3**), and 1.85 µL dH_2O. Reactions were incubated at 37 ºC for 30 mins. The 15 µl A-tailing reaction product was added to 22.4 µl ligation mix, which consisted of 2.24 µl 10x NEBuffer 4, 3.74 µl 10 mM ATP, 0.33 µl 15 µM xGen Duplex Seq Adapters (IDT: 1080799), 0.56 µl 400 U/µl T4 DNA ligase (New England Biolabs), and 15.53 µl dH_2O. Reactions were incubated at 20 ºC for 20 min and subsequently purified with 1x AMPure XP beads and resuspended in 50 µl of nuclease free water.

DNA quantification, dilution and PCR amplification

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

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

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

Library dilution and sequencing efficiency

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

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

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

Using the above equation, we can optimise the efficiency of NanoSeq independently of the input amount of DNA in a given sample. For example, ~0.3 fmols of library yield optimal duplicate rates when using 150 million 150 bp paired-end reads, which are the equivalent of ~15x coverage in standard human whole-genome sequencing. ~0.6 fmol yield optimal efficiency when using 300 million reads (30x whole-genome equivalent). Note that, as predicted by the equations above, deviations ~2-fold from r_{opt} still yield high efficiency. Using these equations we reliably obtained near-optimal duplicate rates from a wide diversity of samples (**Extended Data Fig 4, Supplementary Table 2**). Overall, we found that ~30x of standard sequencing output (~300x10^6 150bp PE reads) yielded approximately 3 Gb of high-accuracy duplex coverage (a haploid genome equivalent) after application of all computational filters.

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

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

Sequencing, preprocessing and filtering of BotSeqS and NanoSeq libraries

Standard sequencing matched-normal libraries were aligned to the human reference genome (GRCh37, hs37d5 build) using BWA-MEM v0.7.5a-r405 [41] with default parameters. Alignments were sorted by coordinate and read duplicates were marked using biobambam2 [42] v2.076 bamsormadup. Matched-normal reads were filtered if marked as duplicate, supplementary, QC fail, unmapped or secondary alignments. For some samples, as described above, instead of standard whole-genome sequencing, we used undiluted NanoSeq libraries (typically ~5 fmol) as matched normals, reducing the costs of sequencing matched normal samples.

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

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

Consensus base quality scores

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

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

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

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

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 $j = 1...d$:

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

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:

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

where

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

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:

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

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

Base calling and filtering

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.

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.

The filters applied are the following:

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

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

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

Indel calling

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

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

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

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

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

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

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

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

Substitution imbalances

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

Genome masks

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

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

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

Detection of human DNA contamination

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

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

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

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 ~$5x10^{-6}$ 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 $5x10^{-6}$ to $9x10^{-8}$ 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 VerifyBamID2 [46], which we evaluated simulating contamination fractions below 1%, for both bams sequenced with standard methods and with the NanoSeq protocol (**Extended Dat 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 regions[47]. 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

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

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

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

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

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

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

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

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

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

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

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

Mutation calling in clonal samples sequenced with standard protocols

Mutation calls for HSPC colonies from donor PD43976_59yo were obtained from Lee-Six *et al*. 2018[21]. Mutation calls from standard whole-genome sequencing for the colonic crypts processed in Lee-Six *et al*. 2019[6] were obtained from Olaffson *et al*.[39]. Indel mutation calls for a bladder tumour sample (**Extended Data Fig 5**) were obtained from Lawson *et al*.[34]. Indel calls for POLE and POLD1 mutants were obtained from Robinson *et al*.[48] (**Extended Data Fig 5**).

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

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

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

Tree building was performed using MPBoot, which is a maximum parsimony tree approximation method[51]. Variants were genotyped as 'present' in a sample if 2 or more variant reads supported the variant. Variants were genotyped as 'absent' in a sample if 0 variant reads were present at a given site and depth at that site was 6 or more. Sites that did not fall into either of the above categories were marked as 'unknown'. Mutations were assigned back to the tree using an R package (tree_mut), which uses a maximum likelihood approach and the original count data to assign each mutation to a branch in the MPBoot generated tree.

Estimation of mutation burden in standard sequencing data

Using clonal or nearly-clonal samples, we were able to compare NanoSeq to mutation burden estimates from standard whole-genome sequencing. This includes libraries prepared by laser microdissection and low-input enzymatic fragmentation[38] or sonication, followed by standard

Illumina sequencing and mutation calling using CaVEMan[49]. The mutation calls described in the previous section were further processed to make burden estimates comparable across protocols.

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

Bootstrapped cosine similarity

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

Mutational signature analysis

Mutational signatures of single-base substitutions in their trinucleotide sequence context were inferred from sets of somatic mutation counts using the sigfit (v2.0) package for R [52]. *De novo* signature extraction was performed for a range of numbers of signatures (N = 2, ..., 8), using counts of mutations grouped per tissue type (cord blood, adult blood, granulocytes, colonic crypts, smooth muscle or neurons), and sequencing method (NanoSeq or standard sequencing). To account for differences in sequence composition across samples, NanoSeq mutation counts were corrected as described in a previous section (see **Correction of mutation burden and trinucleotide substitution profiles**). To avoid an excessive influence of tissue types more highly represented in our dataset, mutation counts were randomly downsampled to a maximum of 2,000 mutations from each tissue type. Samples with evidence of sporadic mutational processes, such as APOBEC or colibactin were removed from the dataset. This excluded urothelium, a bladder tumour sample and colonic crypts from one donor affected by colibactin (PD37449, n = 3). The best-supported number of signatures on the basis of overall goodness-of-fit, as reported by the 'extract_signatures' function in sigfit, was N = 3. The three extracted signatures (**Fig. 3e**) were subsequently fitted to the counts of mutations per sample (using the 'fit_signatures' function in sigfit) to infer the exposure of each signature in each sample.

Mutational signature analysis was also applied to publicly-available single-nuclei mutation data from neurons[13]. Three signatures closely matching those shown in the original publication were extracted using the *extract_signatures* function in sigfit, with parameters nsignatures=3, seed=1469 and iter=10000.

Linear regression modelling

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

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

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

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

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

Supplementary References

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

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

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

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

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

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

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

45 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
46 Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res* **30**, 185-194, doi:10.1101/gr.246934.118 (2020).
47 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72, doi:10.1093/nar/gks001 (2012).
48 Robinson, P. S. *et al.* Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. *bioRxiv*, 2020.2006.2023.167668, doi:10.1101/2020.06.23.167668 (2020).
49 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15.10.11-15.10.18, doi:10.1002/cpbi.20 (2016).
50 Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15.17.11-15.17.12, doi:10.1002/0471250953.bi1507s52 (2015).
51 Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol Biol* **18**, 11, doi:10.1186/s12862-018-1131-3 (2018).
52 Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*, 372896, doi:10.1101/372896 (2020).
53 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
54 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513, doi:10.1073/pnas.1208715109 (2012).
55 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.11-11.10.33, doi:10.1002/0471250953.bi1110s43 (2013).
56 Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623-630, doi:10.1038/nmeth.3407 (2015).
57 Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460-4466, doi:10.1182/blood-2010-08-303537 (2011).
58 Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418-426, doi:10.1038/nature25022 (2018).
59 Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449-1454, doi:10.1126/science.aay9333 (2020).
60 Summers, C. *et al.* Neutrophil kinetics in health and disease. *Trends Immunol* **31**, 318-324, doi:10.1016/j.it.2010.05.006 (2010).
61 Abkowitz, J. L., Catlin, S. N. & Guttorp, P. Evidence that hematopoiesis may be a stochastic process in vivo. *Nat Med* **2**, 190-197, doi:10.1038/nm0296-190 (1996).
62 Abkowitz, J. L., Golinelli, D., Harrison, D. E. & Guttorp, P. In vivo kinetics of murine hemopoietic stem cells. *Blood* **96**, 3399-3405 (2000).
63 Derényi, I. & Szöllősi, G. J. Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations. *Nat Commun* **8**, 14545, doi:10.1038/ncomms14545 (2017).

Data Availability

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

Code Availability

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

Acknowledgements

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

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

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

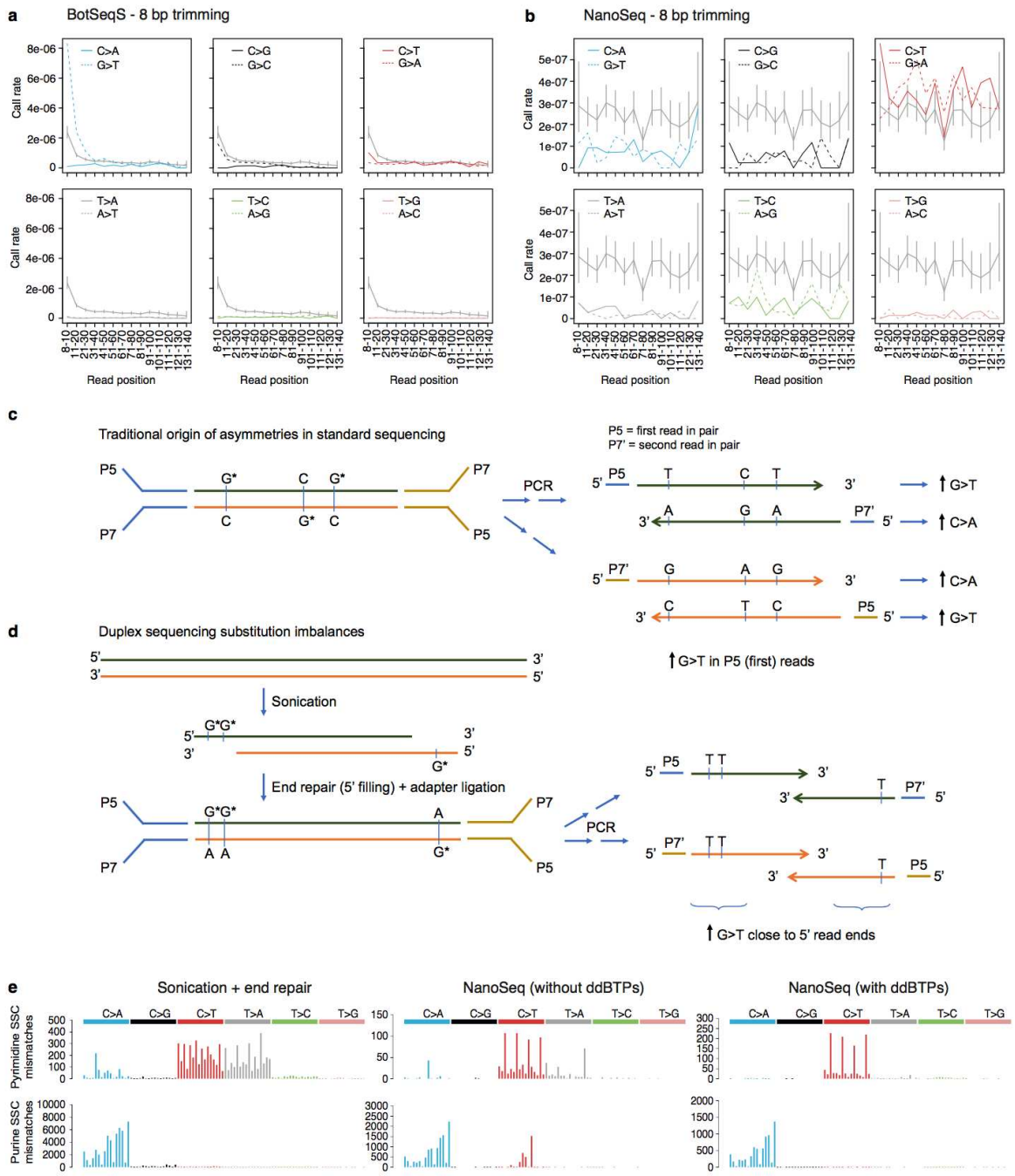
Author Contributions

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

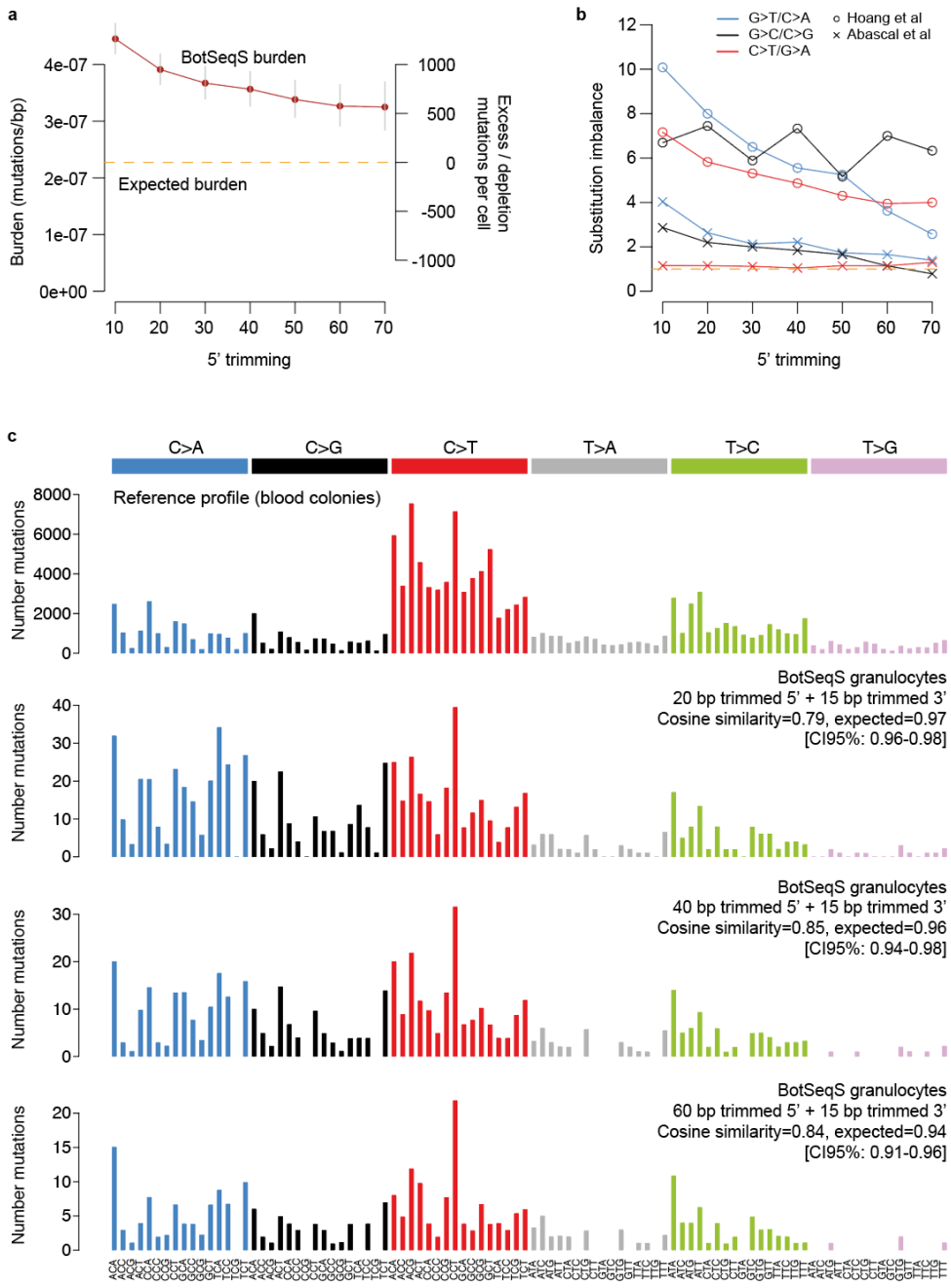
S.V.L. and I.M. R.J.O. and F.A. developed the bioinformatics pipeline with help from R.E.A., S.V.L., and D.J. F.A. led the analysis of the data with help from A.R.J.L., I.M., A.B-O., Y.W., L.M.R.H., E.J.K., T.H.H.C, M.S.C, and M.G. E.M. performed the HSC/MPP experiments. L.M.R.H. and A.J.C.R. performed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

Competing Interests Declaration

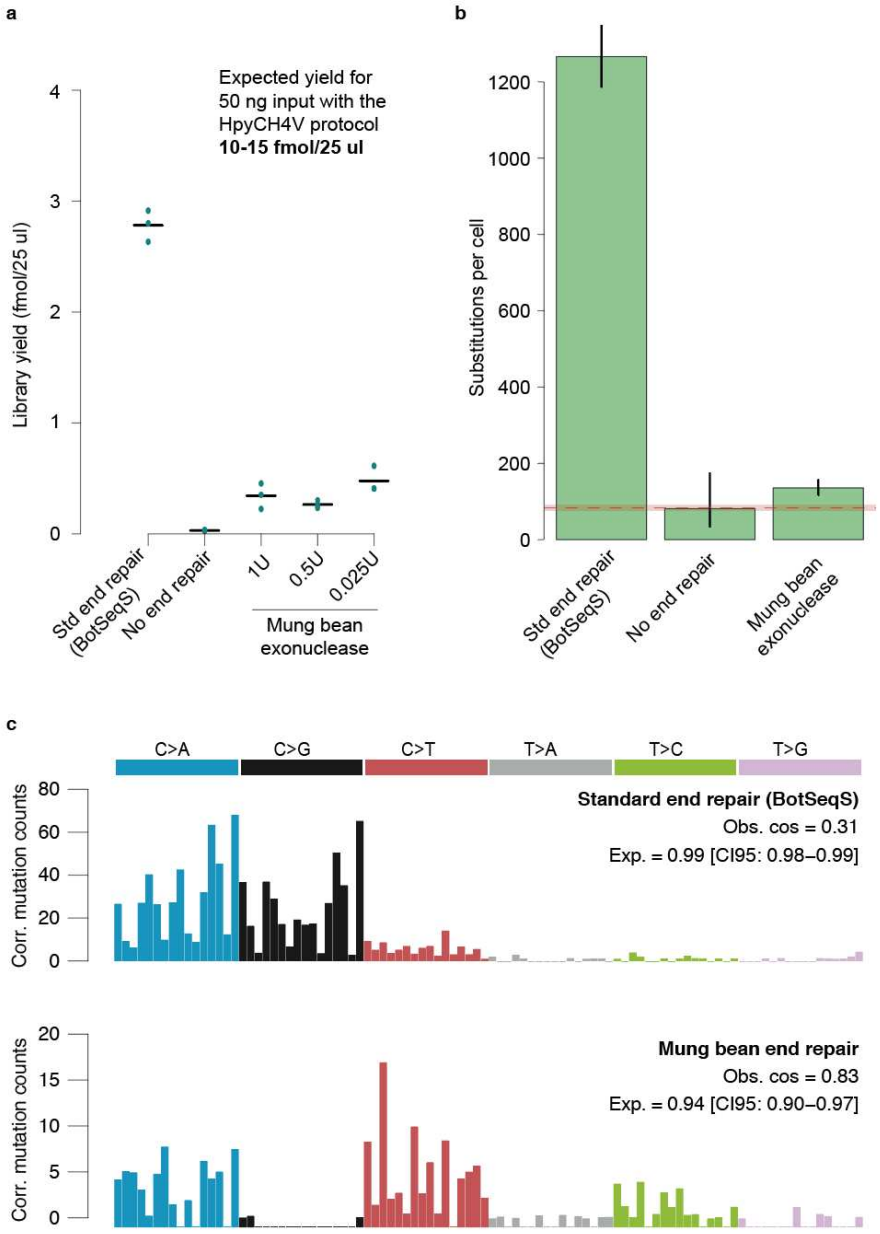
A patent application on NanoSeq is being filed including several authors.



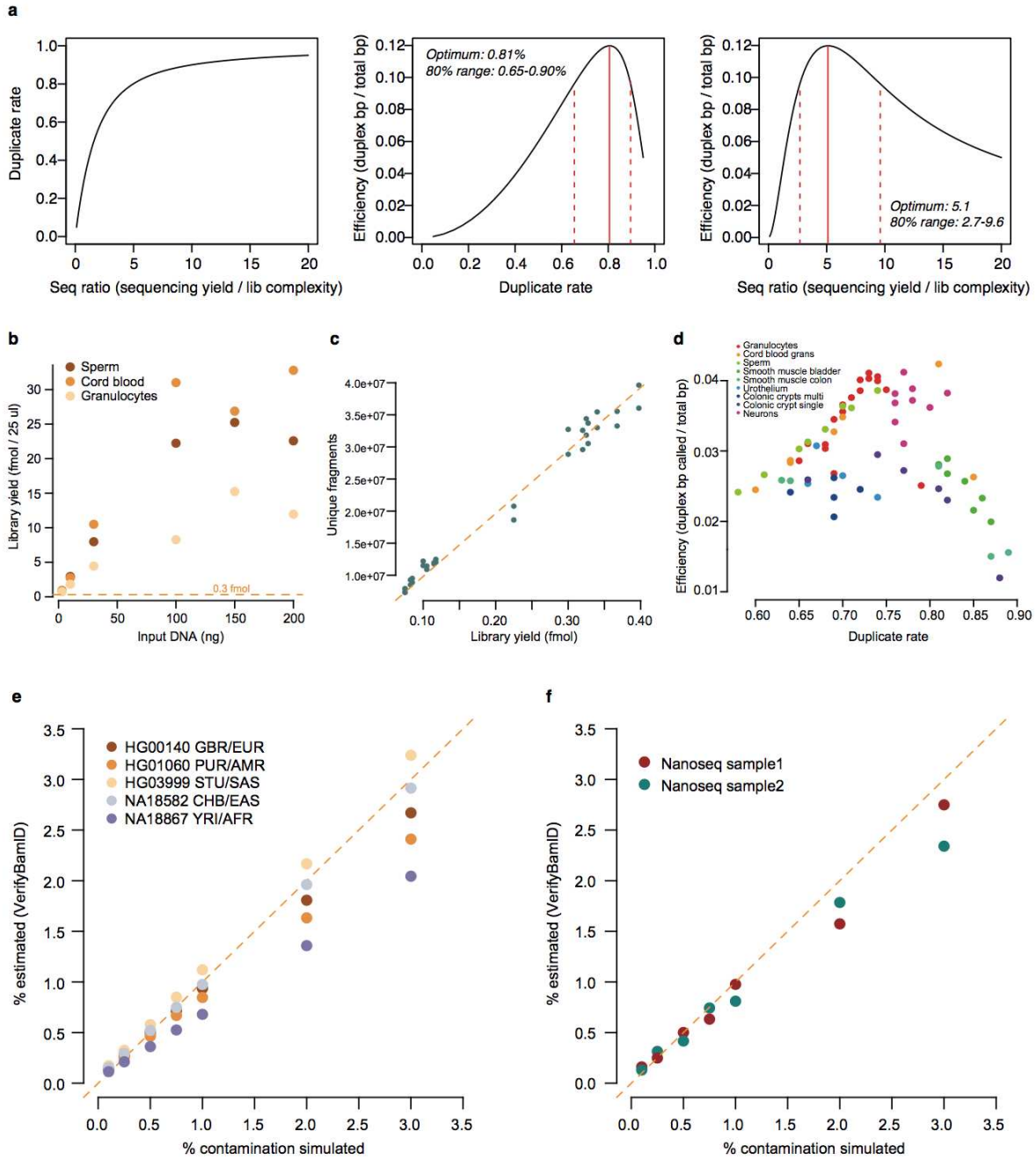
Extended Data Figure 1 | Substitution imbalances and impact of A-tailing. **a-b**, Imbalances in the distribution of the six complementary substitutions (e.g. G>T vs C>A) across read positions in BotSeqS and NanoSeq, respectively. **c**, Origin of G>T over C>A mutation call imbalances in standard sequencing [22]. **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.



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.

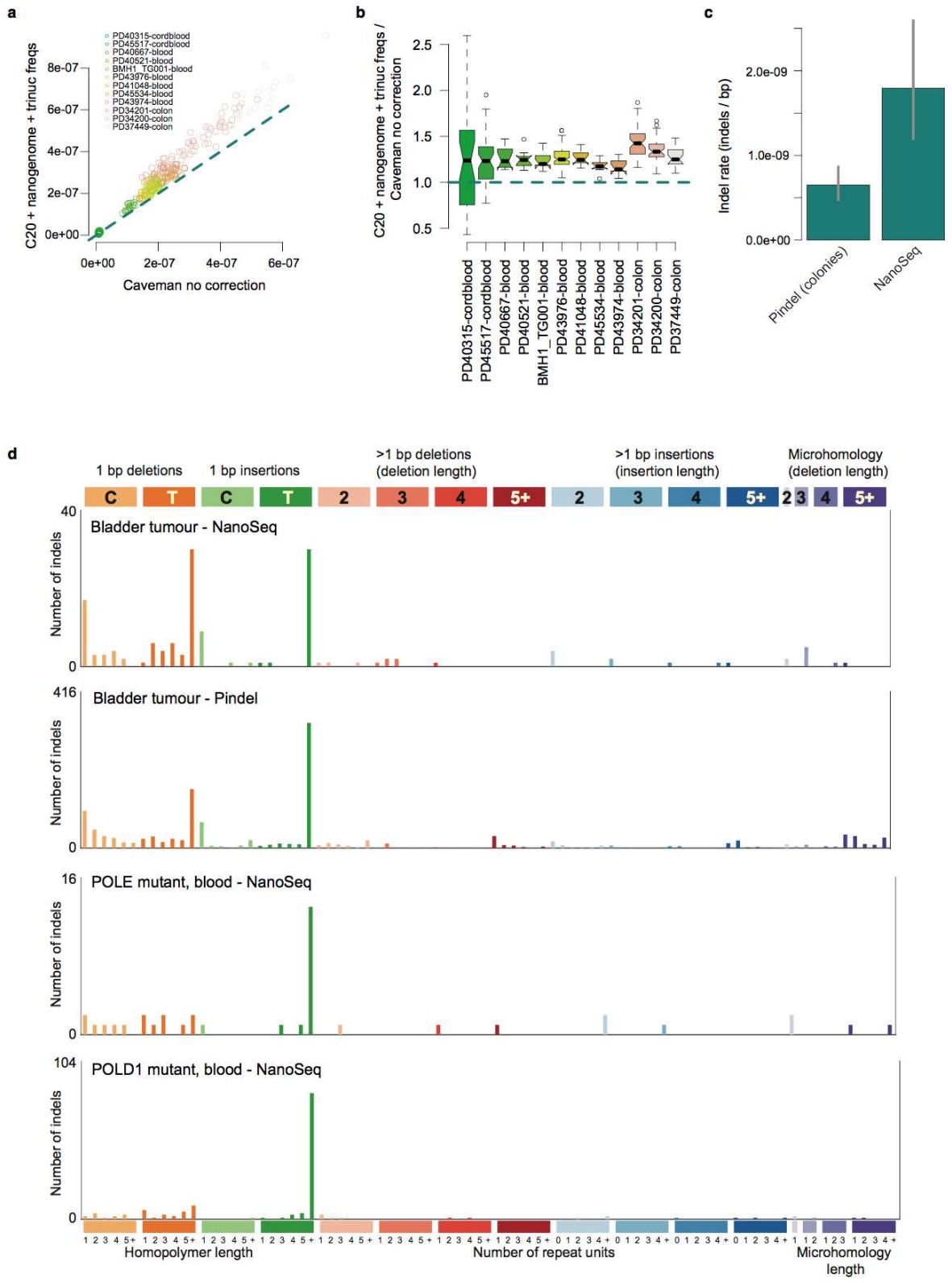


Extended Data Figure 3 | **Alternative protocols for library preparation**. **a**, Library preparation yields for three different kind of protocols run in triplicates (green dots show replicates; mean values as black lines). For Mung bean, different concentrations of the enzyme (U) were tested. **b**, 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.



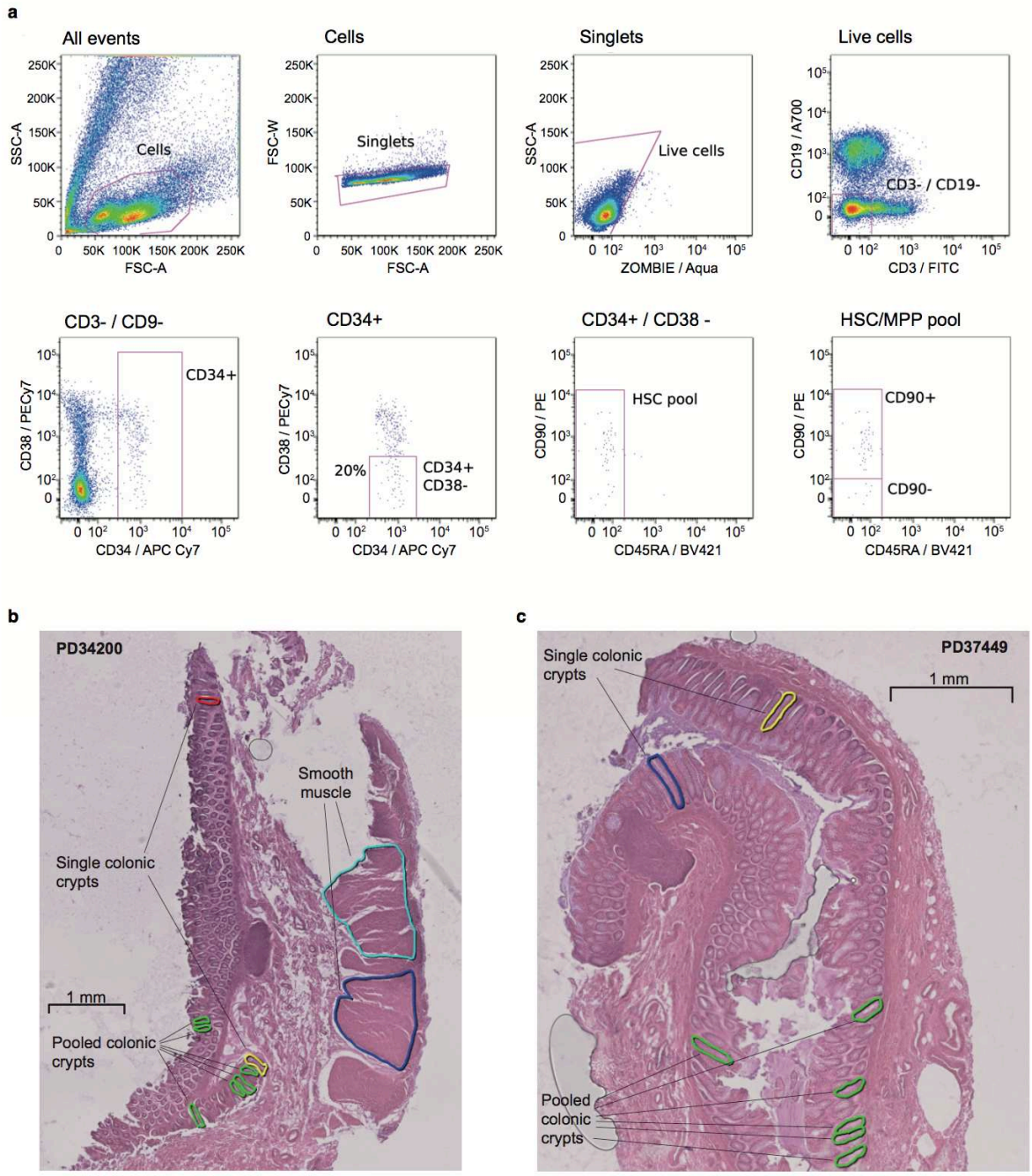
Extended Data Figure 4 | Optimization of duplicate rates, DNA input requirements and estimation of human contamination. **a**, Relationship between sequencing yield, library complexity, duplicate rates and efficiency, based on a truncated Poisson model (**Methods**). From left to right: duplicate rate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

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

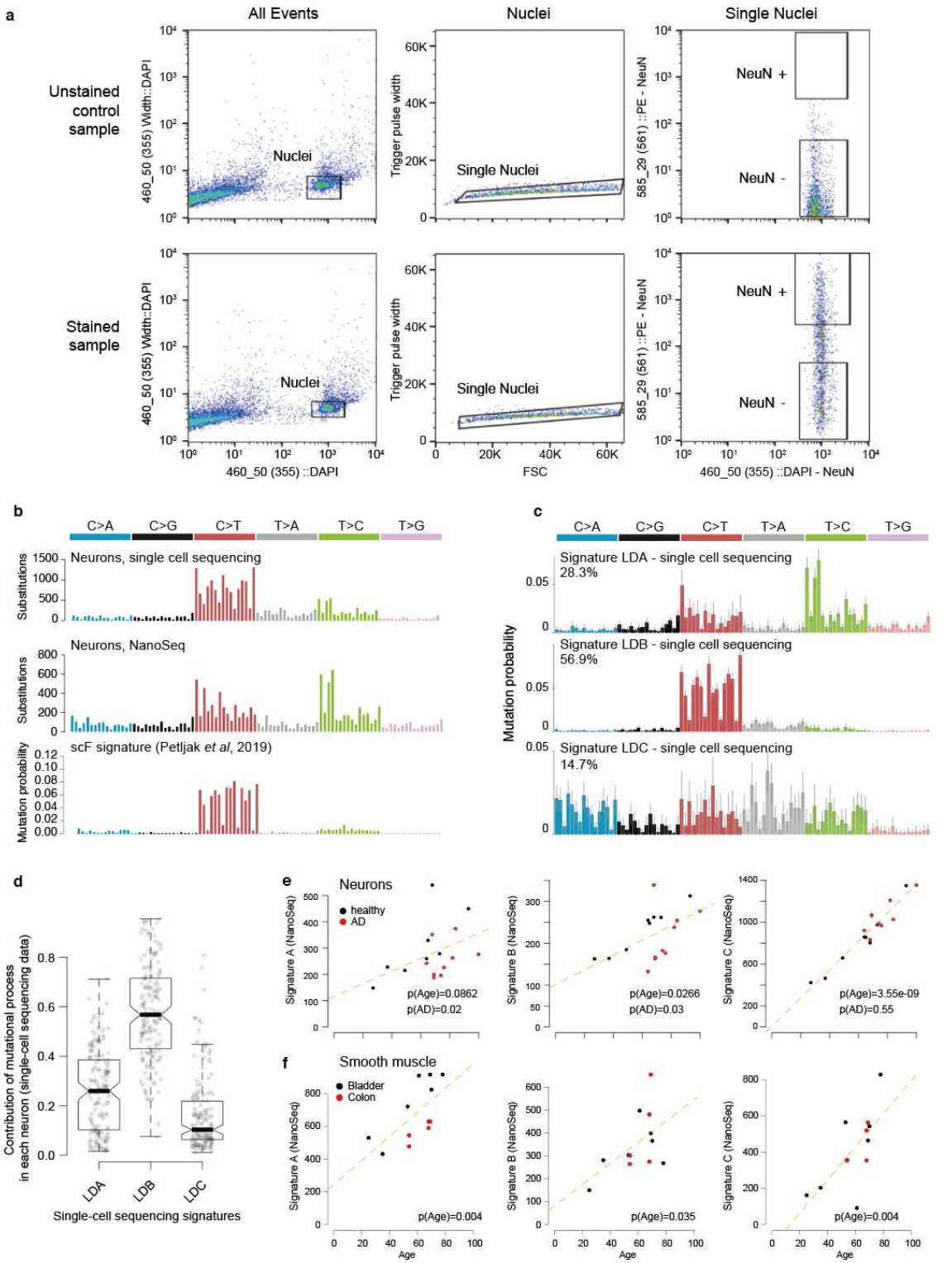


Extended Data Figure 5 | Correction of standard (CaVEMan-based) mutation burden estimates and validation of NanoSeq indel. **a**, Comparison of the mutation burden estimates in regions of the genome with at least 20x coverage (*c*) to the trinucleotide-context-corrected mutation burdens in the subset 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

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

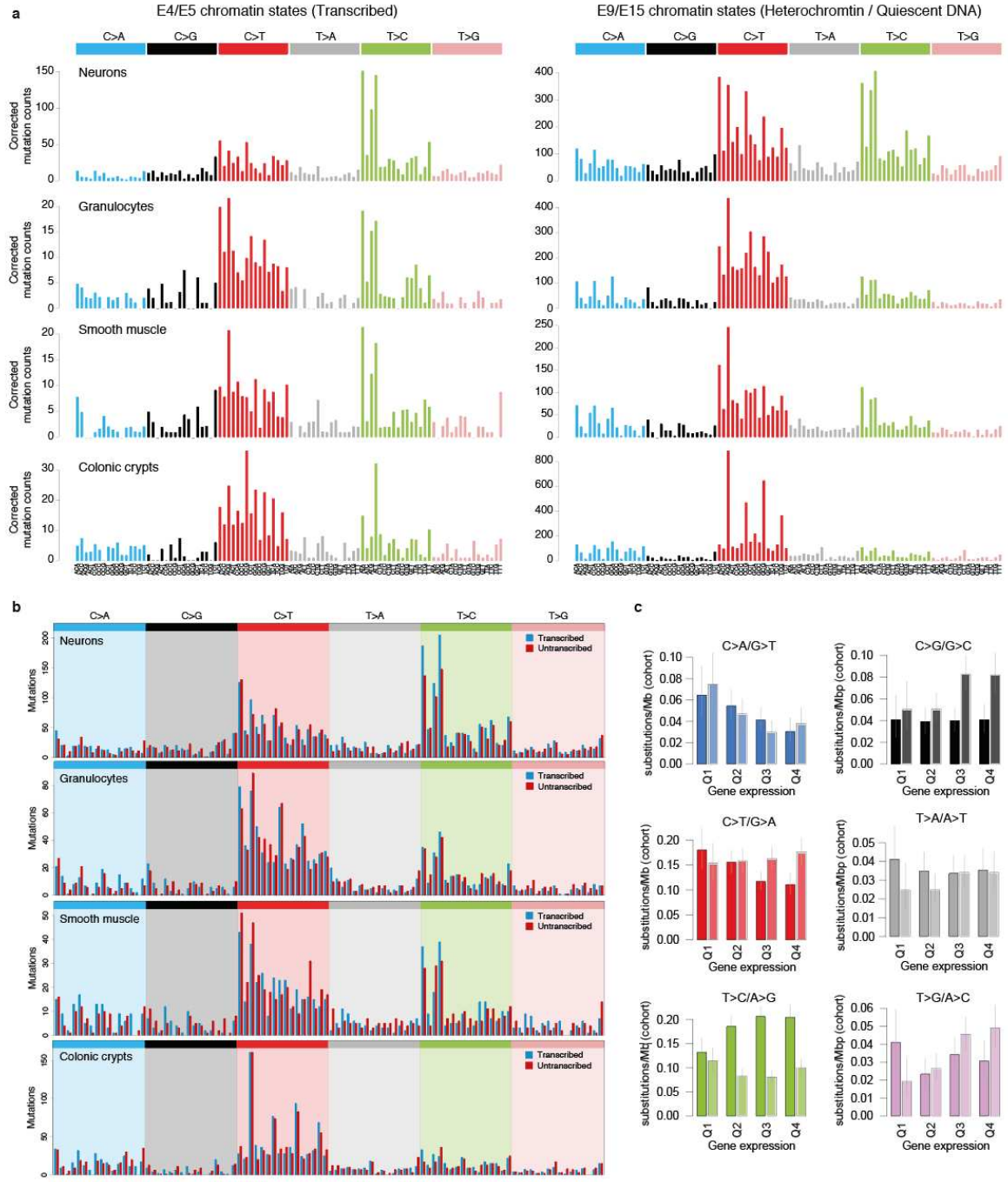


Extended Data Figure 6 | Haematopoietic stem and progenitor cells and colon histology. **a**, Gating strategy for the isolation of HSC/MPPs from a representative bone marrow sample. Text above plots indicates the population depicted. Text inside the plots indicates the name of the gates shown in pink. The CD34+/CD38- population is defined as the bottom 20% CD38- as shown. For all initial samples (BM/PB/CB) the index sorted population is the "HSC pool" gate. Cell population abundance differed between samples but typically viable cells were 60-90% of total cells and singlets were 98-99% of viable cells. Live cells were 90-99% of viable cells and myeloid cells were 15-50% of live cells. CD34+ cells were typically 1-15% of myeloid cells. **b** and **c**, Colon histology sections showing microbiopsied areas of colonic epithelium and smooth muscle for donors PD34200 and PD37449, respectively.

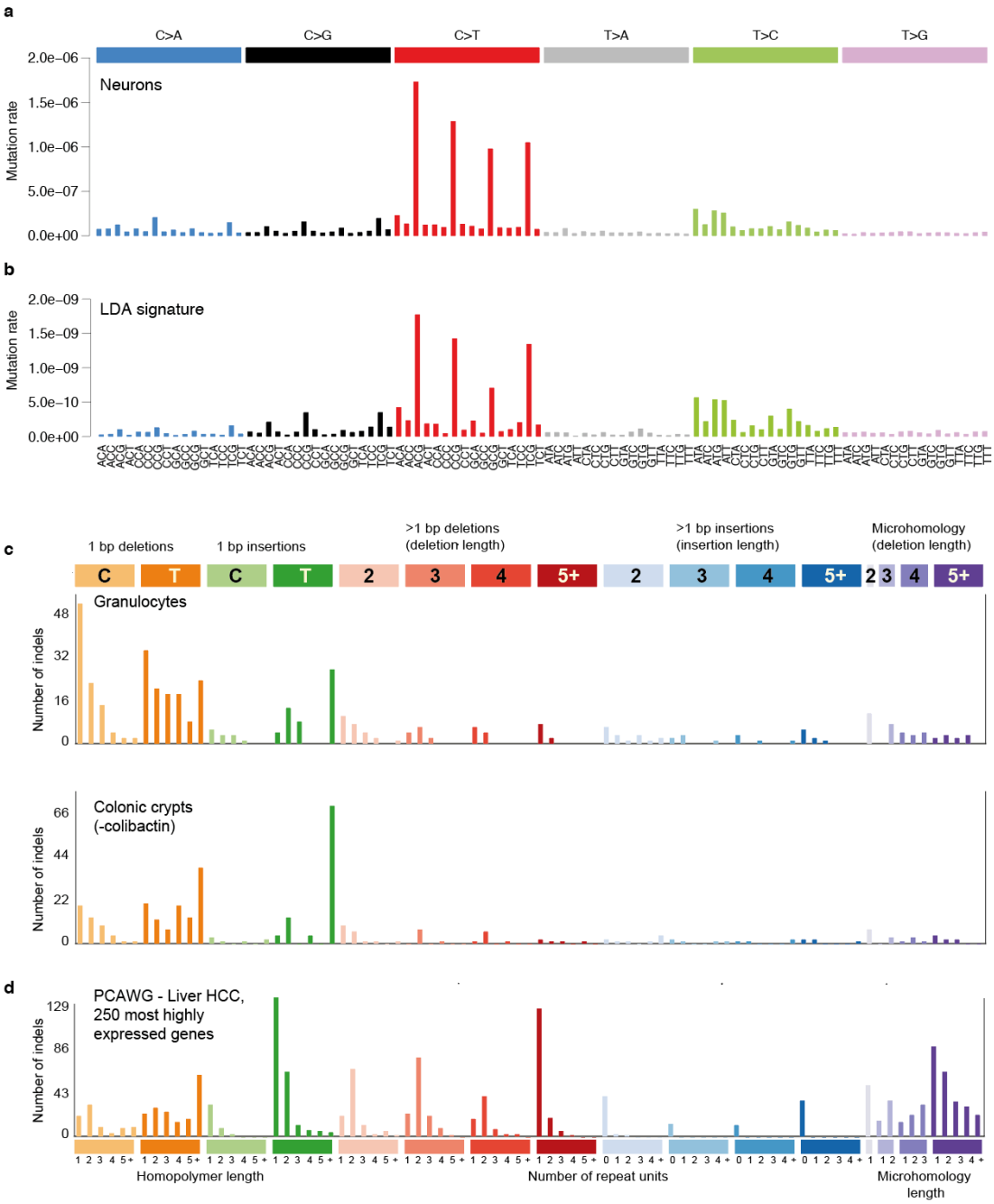


Extended Data Figure 7 | Neuron nuclei sorting, comparison to single-cell data and accumulation of mutations with age. **a**, Gating strategy for the isolation of neuronal nuclei from frontal cortex. Nuclei were sorted by FACS using an Influx cell sorter (BD Biosciences) with a 100-µm nozzle. For each sample an unstained control was used to help determine the NeuN+ population. The text above each column indicates the population depicted and the text inside the plots indicates the population of the gates highlighted in black. Sorting results varied among samples, with 1-60% passing the DAPI gate and, of these, 2-53% passing a conservative

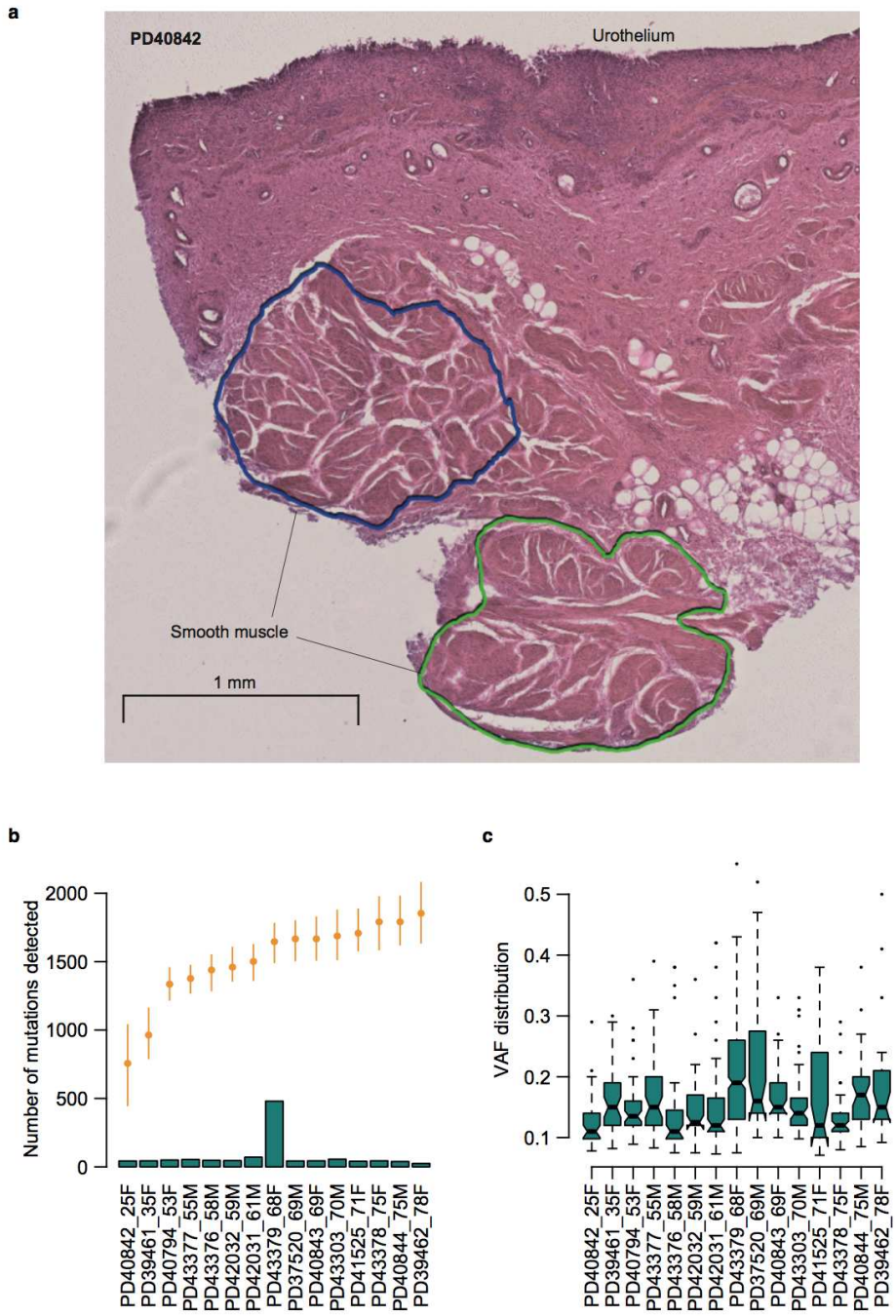
NeuN+ gate. **b**, Substitution profiles for all mutations detected in neurons with SNP-phased error-corrected single-cell sequencing data in Lodato *et al*.[13] (top) and with NanoSeq (middle). In the bottom panel, a signature specific of single-cell sequencing data is shown (scF signature from Petjak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



Extended Data Figure 8 | Normalised substitution spectra across different genomic regions. **a**, Substitution spectra for neurons, granulocytes, smooth muscle and colonic crypts in chromatin states associated to transcription (states E4 and E5 in ENCODE), and inactive DNA (E9 and E15). Chromatin states were obtained from ENCODE[53], using the following epigenomes: E073 (frontal cortex), E030 (granulocytes), E076 (smooth muscle), and E075 (colonic mucosa). To enable direct comparison of spectra across genomic regions with different trinucleotide frequencies, the profiles have been normalised to the genomic trinucleotide frequencies (**Methods**). **b**, Transcriptional strand asymmetries in neurons, granulocytes, smooth muscle and colonic crypts. **c**, Transcriptional strand asymmetries in neurons in quartiles of gene expression.



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 data[31].



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 protocol[38]. 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.

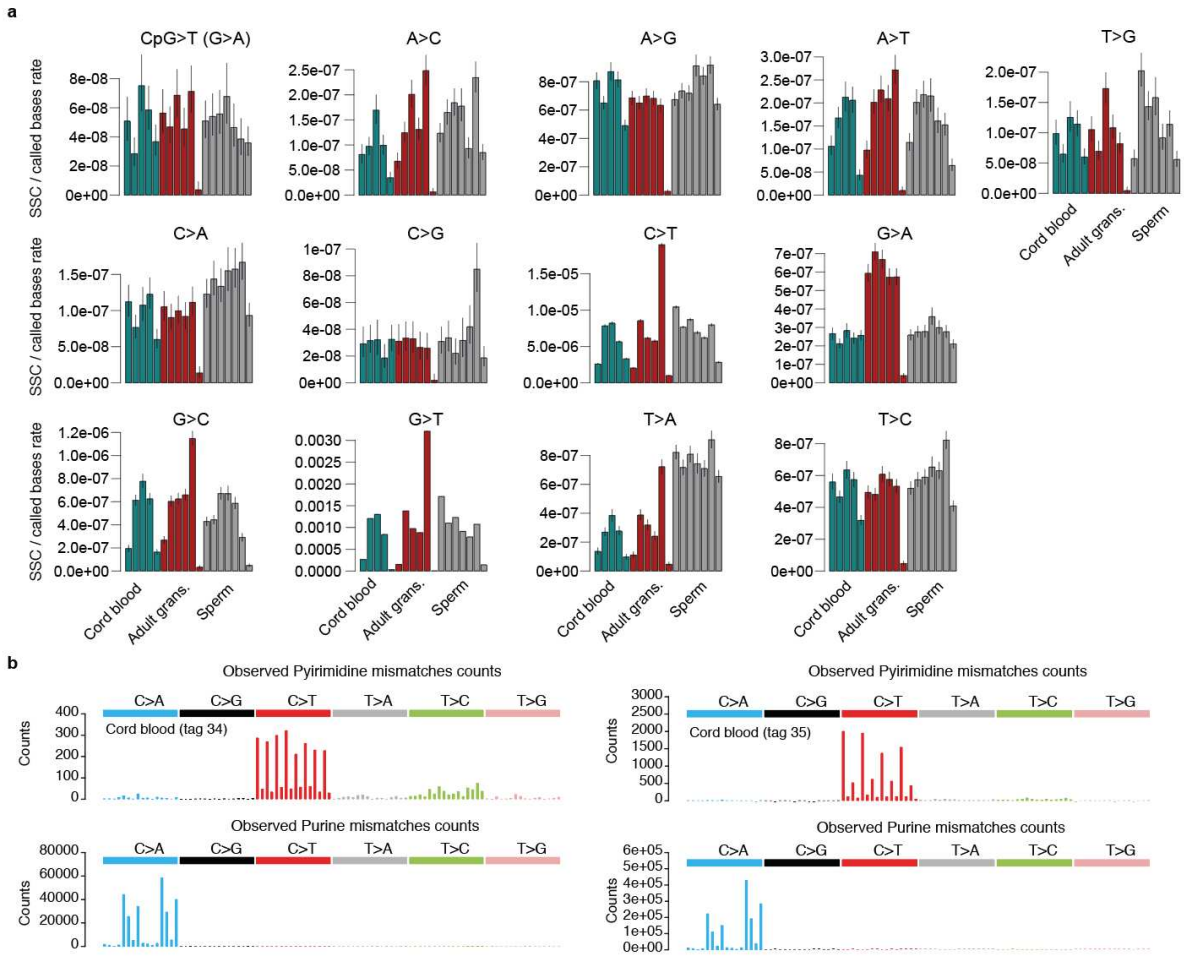
Supplementary Notes

Supplementary Note 1 - Single strand consensus calls

In duplex sequencing protocols, consensus calls from only one of the two strands (single strand consensus calls) can be caused by amplification artefacts or by DNA damage[54], including damage originated in vivo or in vitro, during DNA extraction, storage or library preparation. Given the value of measuring in-vivo DNA damage, we explored whether information on DNA damage could be extracted from the abundant single strand consensus (SSC) calls.

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

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



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.

Supplementary Note 2 - Restriction enzyme choice

We identified 14 commercially-available restriction endonucleases with 4 base-pair recognition sites that generated 5′ overhangs or blunt ends and were not impaired by overlapping CpG methylation. We computationally digested the human genome (hs37d5) with each restriction enzyme and, assuming size selection of fragments between 250 and 500 base pairs and 150 bp paired-end reads, calculated the coverage for the whole genome, the coding genome and the mitochondrial genome (**Supplementary Table 3**). The candidates with the highest coverage included AluI, CviAII, FatI, and HpyCH4V, of which only AluI and HpyCH4V leave blunt ends. All four enzymes have a recognition site with 50% GC content. We opted for HpyCH4V given its higher whole-genome and coding coverages, although its mitochondrial coverage was lower than that of AluI.

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

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

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

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

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

We then compared the mutation burden and mutational spectra across protocols (**Extended Data Fig 3b**). As expected, sonication followed by standard end repair (BotSeqS) yielded a high error rate, with around 1,200 errors per diploid genome (~$2x10^{-7}$ errors/bp) and a mutational spectrum dominated by C>A and C>G errors (**Extended Data Fig 3c**). Sonication followed by Mung Bean nuclease or no end repair yielded low mutation burdens, similar to those using the restriction enzyme protocol, with error rates estimated to be in the nano scale ($<10^{-8}$ errors/bp). Libraries generated without end repair or Mung Bean nuclease did not produce enough library yield to enable a detailed comparison of mutation burdens and spectra. The mutational spectra of the Mung Bean nuclease libraries were largely consistent with that of cord blood single-cell derived colonies, with a cosine similarity within the expected 95% confidence interval (**Methods**), although the rate of C>A mutations appeared to be slightly elevated (**Extended Data Fig 3c**).

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

Supplementary Note 4 - Modified A-tailing

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

restriction sites, and the profile of single-strand consensus, typically caused by DNA damage, showed an increased amount of G>A, C>A and T>A (**Extended Data Fig 1e**). To explain this pattern we hypothesised a multi-step mechanism involving: nicking of the DNA duplex by restriction enzymes (an intermediate step in double-strand cleavage); 3′ to 5′ exonuclease or pyrophosphorolysis, during A-tailing, of the dNTP 3′ of the nick; incorporation of dATP opposite C, G or A during A-tailing (causing G>A, C>A or T>A, respectively); and subsequent sealing of the internal nick during adapter ligation. To block molecules with internal nicks or gaps, we replaced dATP with a mixture of dATP and ddBTP (ddCTP, ddGTP, ddTTP) during A-tailing. The presence of internal nicks would trigger DNA polymerase extension until the incorporation of a ddBTP, making the affected DNA strand unamplifiable. Our results show that the incorporation of ddBTPs successfully removed artefacts caused by A-tailing (**Extended Data Fig 1e**).

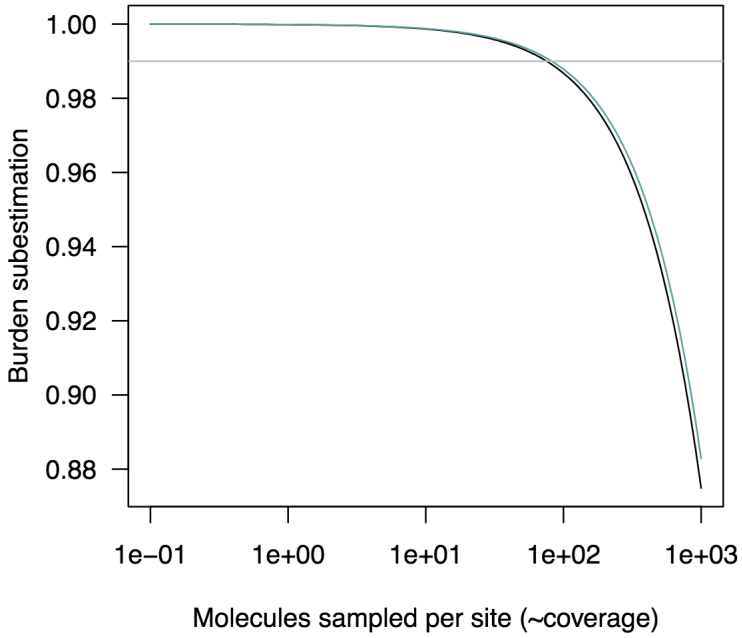
Supplementary Note 5 - Chimeric read bundles

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

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

To test for the presence of chimeric bundles empirically, we can study the fraction of read bundles that contain both alleles of a heterozygous SNP. We focused on donor PD43976 (**Supplementary table 1**), for which multiple colonies are available[21]. We ran GATK's HaplotypeCaller[55] on each of the colonies and detected 1.4 million reliable heterozygote SNPs in the donor, defined as those called in at least 90% of the samples and showing a VAF between 0.4 and 0.6. We estimated how many times these heterozygote SNPs passed mapping quality filters and were seen in two NanoSeq libraries from this donor, and how many times a

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



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

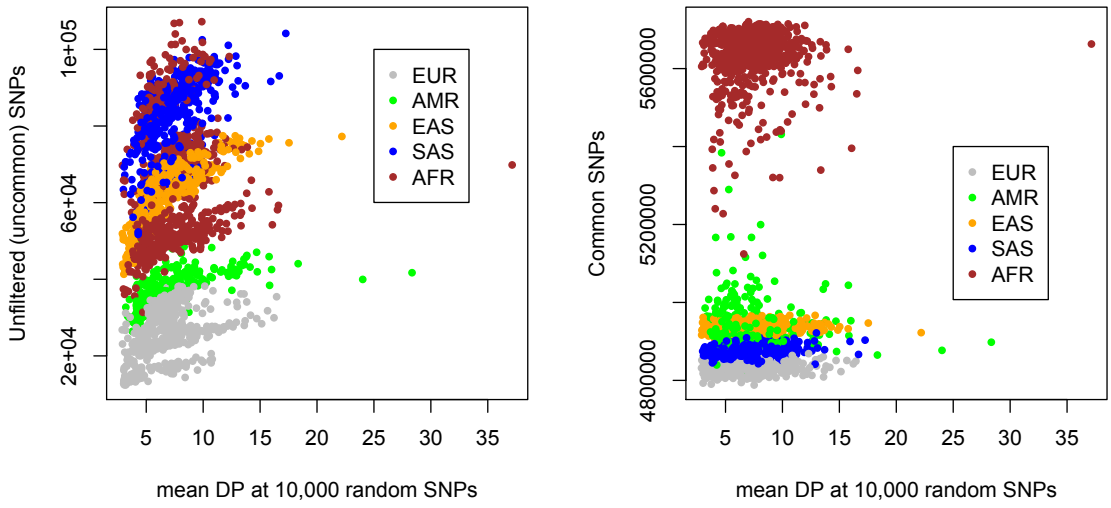
Supplementary Note 6 - Human DNA contamination

To reduce the impact of contamination on duplex sequencing libraries we generated an extensive SNP mask (**Methods**). For each NanoSeq library we also estimate the extent of inter-individual contamination using VerifyBamID2[46], which we validated using simulations of contamination fractions as low as 0.1% (**Extended Data Fig 4e**,**f**).

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

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

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



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

These analyses confirmed that, depending on the mutation burden of the sample and the ancestry of the contaminant, 1% of contamination can still be problematic after application of the SNP mask. VerifyBamID is a tool routinely used to estimate human contamination from sequencing data. The most recent version[46] is ancestry aware and has been tested for contamination levels above 1%. Here we performed simulations to evaluate VerifyBamID performance at 0.1, 0.25, 0.5, 0.75, 1, 2, and 3% contamination levels. To obtain more stable estimates of contamination 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.

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

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

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

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

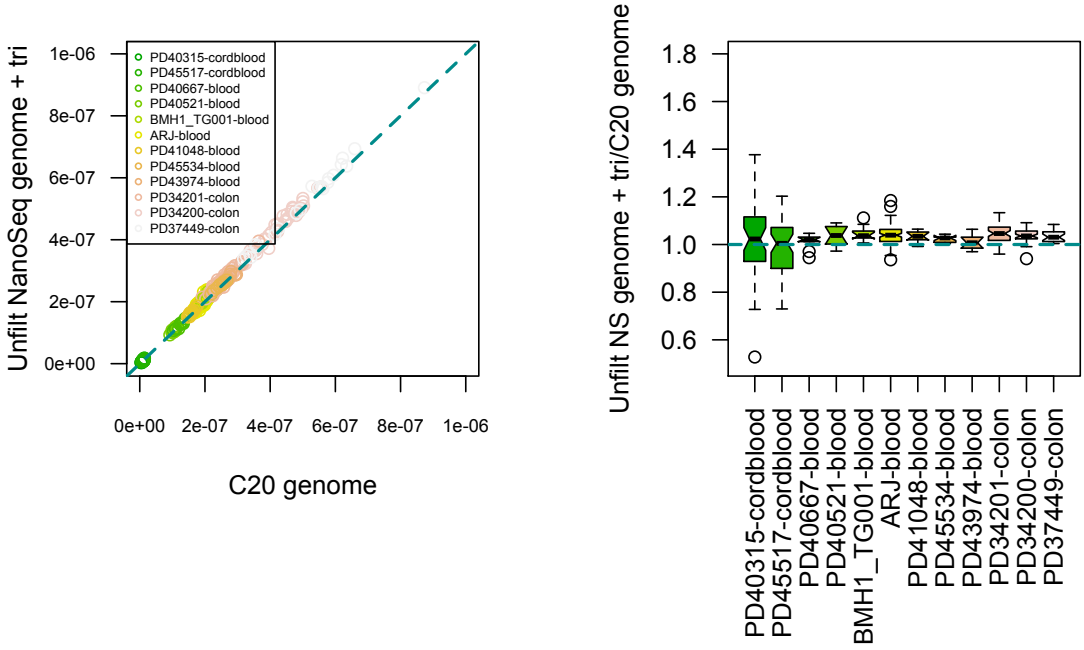
Mutation burden estimation in NanoSeq is unaffected by the clonality of a sample or the depth of coverage. In contrast, the somatic mutation calling sensitivity in standard sequencing data depends on both clonality and coverage. Except for standard sequencing of smooth muscle (for which we did not attempt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

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

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

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.



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.

Supplementary Note 8 - Validation of indel calls

To estimate the indel error rate we compared Pindel calls[50] in single-cell derived cord blood colonies to our indel calls in NanoSeq cord blood granulocytes. Comparing indel rates is particularly difficult given the known problems of specificity and sensitivity associated with indel calling. For this comparison we applied the same approach that we used to compare CaVEMan and NanoSeq, restricting the analysis to regions with at least 20x coverage and falling in the NanoSeq-covered genome. Pindel estimated 6.5 x 10^{-10} indels / bp (CI95% 4.8 x 10^{-10} - 8.7 x 10^{-10}), while NanoSeq estimated 1.8 x 10^{-9} indels / bp (CI95% 1.2 x 10^{-9} - 2.6 x 10^{-9}; **Extended Data Fig 5c**). Although NanoSeq estimates are higher, and some of this difference may be due to higher rates of indels in differentiated cells, we can confidently estimate an indel error rate for NanoSeq <3 x 10^{-9} / bp.

The reliability of our indel calls is further supported by the linear accumulation of indels with age observed for granulocytes, smooth muscle and neurons. To further investigate the quality of our indel calls we also compared indel profiles for samples with reliable indel calls from standard whole-genome sequencing data. These included a bladder tumour, colonic crypts (with and without the colibactin signature), and *POLE* and *POLD1* mutants. NanoSeq indel profiles matched closely the reported indel profiles for colibactin[27], POLE/POLD1 samples[48], and a previously published bladder tumour sample[34] (**Extended Data Fig 5d**).

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

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

Estimates of the population size of human haematopoietic stem cells (HSC) range from 25,000 to 1.3 million and the human HSC division rate is estimated to be between one per 28 days and one per 4 years[21,57-59]. The relatively small population size and division rates of HSC contrast with the staggering production of blood cells throughout life. On the order of 10^{11} granulocytes are estimated to be produced every day[60] and on the order of 1.4 x 10^{14} blood cells are estimated to be produced every year considering all mature cell types[58].

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

Theoretically, the minimum number of divisions required to produce N differentiated cells from a single cell is achieved by a perfectly bifurcating tree[63], in which a single cell expands into $N = 2^d$ cells. Hence, the minimum number of cell divisions that must separate a HSC from an average differentiated cell can be calculated as $d = \log_2 N$. If we assume, as an example, that the HSC pool in humans is 100,000 cells and that 10^{11} granulocytes are produced every day, then the minimum number of cell divisions with a perfect bifurcating tree would be $d = \log_2(10^{11}/10^5)$, that is, at least 19.9 cell divisions. However, we know that HSC divide infrequently (around once per year) and need to self-sustain. To maintain homeostasis, on average, the division of a HSC results in a HSC and a progenitor. Because progenitors have to

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

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

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

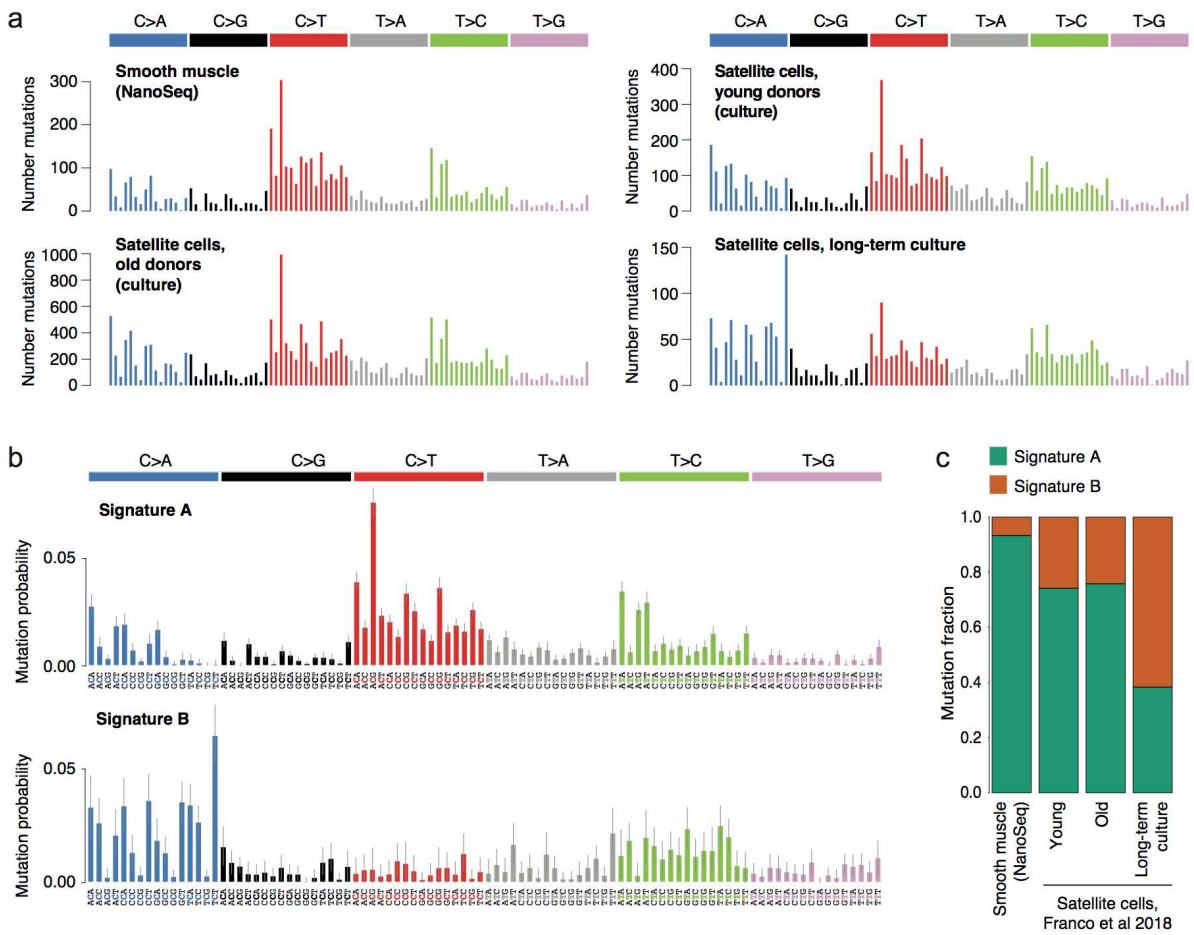
Supplementary Note 9 - Comparison between smooth muscle and single-cell derived colonies of skeletal muscle satellite cells

To our knowledge, our manuscript contains the first description of the mutational landscape of smooth muscle. A previous study described the mutation rates and mutational spectrum of skeletal muscle satellite cells by expanding single satellite cells into colonies in vitro[11]. In that study, the authors sequenced colonies from young and old donors, as well as colonies grown in vitro for different lengths of time, to quantify the effects of in vitro culture.

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

Using sigfit we extracted two signatures from the set of four substitution profiles composed of NanoSeq smooth muscle, satellite cells from young and old donors, and long-term cultured colonies (**Supplementary figure 5b**). Signature A is very similar to NanoSeq smooth muscle (cosine of 0.99), whereas signature B is similar (cosine of 0.84) to signature C in Blokzijl et al[10], which is associated to mutations introduced during in vitro culture. This signature has a

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



Supplementary figure 5. **a**, Substitution profiles of smooth muscle (top left) and satellite cells from young and old donors, and in long-term culture[11]. **b**, Two extracted signatures from satellite cells. **c**, Estimated exposure of groups of samples to each of the two extracted signatures, showing how signature B contribution becomes stronger in long-term culture and is practically absent from NanoSeq data.