Somatic mutation landscapes at single-molecule resolution

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 T. Mahubani[4,5,6], 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

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 with error rates <5 errors per billion base pairs in single DNA molecules from cell populations. This rate is two orders of magnitude lower than typical somatic mutation loads, enabling the study of somatic mutations in any tissue independently of clonality. We exploited this single-molecule sensitivity 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 considerably more divisions. We then characterised the mutational landscape of post-mitotic neurons and polyclonal smooth muscle, confirming that neurons accumulate somatic mutations at a constant rate throughout life without cell division, with similar rates to mitotically-active tissues. Altogether, our results suggest

that mutational processes independent of cell division are important contributors to somatic mutagenesis. We anticipate that the ability to reliably detect mutations in single DNA molecules could transform our understanding of somatic mutagenesis and enable non-invasive studies on large-scale cohorts.

Introduction

Somatic mutations occur in our cells as we age. Because most somatic mutations are present in small groups of cells or even in single cells, studying somatic mutagenesis has been challenging, requiring special approaches. This includes 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], and single-cell sequencing[12-14]. While these technologies are changing our understanding of somatic mutagenesis, the error rates of single-cell approaches have, until recently[15], been too high[16], and other approaches are limited to mitotically-active cell types.

As a result of these 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 differentiated cells that make the bulk of mitotically-active tissues and are responsible for tissue function, and 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 inform on the contribution of cell division and DNA replication to somatic mutation in human tissues. To address these questions, here we present a new sequencing protocol that enables the study of somatic mutations in any tissue or cell population by reliably detecting somatic mutations in single DNA molecules.

Nanorate sequencing

Several protocols have been developed to increase the accuracy of standard sequencing by barcoding individual molecules of DNA and sequencing each molecule multiple times, reducing error rates by single-molecule consensus[17]. The most accurate approaches use duplex consensus sequencing[18,19], 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, the probability of two early and complementary PCR errors in both strands[17]. Given that this rate is lower than the typical mutational load of human tissues, it raises the possibility of quantifying somatic mutation rates in genetically-heterogeneous samples, by detecting somatic mutations on single DNA molecules. This is the rationale of BotSeqS, a whole-genome duplex sequencing protocol[20] (**Fig. 1a**). In practice, however, mapping errors and some library preparation artefacts can violate the assumed independence of both strands[20,21]. The actual error rates of duplex sequencing protocols have remained difficult to measure due to the lack of control samples with low and known mutation rates[17].

To evaluate the performance of BotSeqS, we used samples of cord blood, comparing BotSeqS on bulk granulocytes from a neonate to standard sequencing of 100 single-cell-derived colonies from two neonates as a control. On average, single-cell-derived colonies had 66 mutations per cell, dominated by C>T mutations at CpG sites. In contrast, BotSeqS estimated 1,240 mutations per diploid genome, dominated by C>A and C>G (**Fig. 1b**, **c**). Analysing the distribution of substitutions across BotSeqS reads revealed a large excess of

G>T/C and C>T substitutions near the 5' ends of DNA fragments, and an imbalance over the complementary C>A/G and G>A substitutions affecting the entire read length (**Fig. 1d**, **Extended Data Figs. 1** and **2**). These imbalances are incompatible with real mutations and reflect errors introduced during library preparation[22] (**Methods**, **Supplementary Note 1**). We found the same imbalances, with a much larger C>T component, in the original BotSeqS publication[20] (**Fig. 1d**). Extensive trimming of read ends only partially alleviated these errors (**Extended Data Fig. 2**). Overall, we estimate that BotSeqS introduced approximately 1,200 errors per diploid genome in our samples (i.e. ~2 x 10^{-7} errors/bp).

Based on the error patterns, we reasoned that end repair was likely responsible for most errors, by converting DNA damage in single-strands of DNA into double-stranded errors (**Fig. 1e** and **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**, **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. They also 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 non-A dideoxynucleotides (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 create single-molecule-derived read families (**Supplementary Note 5**).

If duplicate rates are not optimised, duplex sequencing approaches can suffer from low efficiency due to suboptimal read family sizes[20]. We use mathematical modelling of family sizes and qPCR quantification of the library to maximise the duplex coverage independently of the amount of input DNA (**Methods**, **Extended Data Fig. 4a-d)**. A robust bioinformatic pipeline was also developed to avoid false positive mutation calls from mapping errors and from low-level DNA contamination (**Extended Data Fig. 4e**, **f**, **Methods**, **Supplementary Note 6**), and to distinguish germline from somatic mutations.

Applying NanoSeq to cord blood granulocytes yielded an estimated mutation rate of 109 mutations per cell (95% Poisson confidence intervals 95-125; **Fig. 1g)**. The small difference with the colonies could be due to NanoSeq errors, a higher mutation burden in granulocytes than in cord blood stem cells, or both. Consistent with most mutations detected by NanoSeq being genuine, no substitution imbalances were detected in the NanoSeq calls (**Fig. 1d**) and no significant differences were found between the mutational spectra of colonies and granulocytes (**Fig. 1c**, **Methods**). As an additional low-burden control, we applied NanoSeq to a sperm sample from a 21-year-old donor. Seven NanoSeq replicates of the sperm sample yielded low mutation burdens, with ~52 mutations per haploid sperm cell (1.8 x 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**). Together, the sperm and cord blood data indicate that the error rate of NanoSeq is lower than 5 x 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) also revealed an indel error rate <3 x 10^{-9} errors/bp (**Methods**, **Extended Data Fig. 5c**, **Supplementary Note 8**).

The extremely low error rate of NanoSeq, in the nano range, enables the reliable detection of somatic mutations in single DNA molecules, opening the door to the study of somatic mutations in any tissue or cell population. We take advantage of this unprecedented ability to 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.

Mutation burden in stem and differentiated cells

Due to technical limitations, most of our knowledge of somatic mutagenesis is restricted to stem or proliferating cells. Since stem cells are believed to be better protected against mutations[25], differentiated cells could conceivably have higher mutational loads and undescribed mutational signatures[14].

We first addressed this question in the haematopoietic system, comparing mature granulocytes to haematopoietic stem and multipotent progenitor cells (HSC/MPPs) (**Methods**). The haematopoietic system is organised hierarchically, with a heterogeneous pool of slow-cycling stem cells sustaining the production of large numbers of differentiated cells through the extensive proliferation of intermediate progenitor cells (**Fig. 2a**). HSCs are estimated to divide around once a year and conservative estimates suggest that an average of over 28 cell divisions must separate stem cells from differentiated cells to explain the production of ~10^{14} mature cells per year (**Fig. 2a**, **Supplementary Note 9**). As a result, a considerably higher mutation burden and mutational signatures associated with the proliferation of progenitors may be expected in granulocytes.

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[26] (**Methods**). These data revealed remarkably similar mutation burdens in terminally-differentiated granulocytes and HSC/MPPs (**Fig. 2b**). Linear mixed-effect regression yielded indistinguishable slopes for HSC/MPPs colonies and granulocytes (P=0.92), with a joint estimate of ~19.9 mutations/year (CI95% 18.3-21.4, **Methods**, **Supplementary Table 8**). The excess of mutations in granulocytes over HSC/MPPs was estimated to be ~51 mutations and not significantly different from zero (CI95%: -14-120, P=0.13, **Methods**, **Supplementary Table 8**). Their mutational spectra were also largely similar (cosine similarity 0.98, **Fig. 2c**).

The observation that a considerable increase in cell divisions does not cause a proportional increase in mutation burden suggests that replication errors cannot be responsible for more than a small minority of mutations in HSC/MPPs (**Supplementary Note 9**). A caveat for this comparison is that HSC/MPP colonies successfully grown in vitro may not reflect the mutation rate of the more quiescent HSCs responsible for long-term maintenance of the haematopoietic system. However, a similar conclusion can be drawn from the granulocyte data alone. The strong linear relationship with age and the small intercept for granulocytes alone (142.1 mutations, CI95%: -115.3-414.2, compared to the slope of ~19.8 mutations/year) suggests that the majority of the mutations observed in adult granulocytes accumulated in the 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**).

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 single Lgr5+ cells[10] and from sequencing single laser-microdissected colonic crypts[6], which over time become clonally derived from a single stem cell[27]. For three previously-studied donors we compared standard whole-genome sequencing of microdissected colonic crypts[6] to NanoSeq data from single crypts or groups of crypts (**Extended Data Fig. 6b**, **c**). This revealed similar estimates of mutation burden, despite the time lag to clonality in standard sequencing of colonic crypts (**Fig. 2d**). Mutation burden and signatures from differentiated cells in colonic epithelium were consistent with those found by previous studies on colonic stem cells, with a dominance of SBS1, SBS5 and, in some donors, a colibactin signature[28] (**Fig. 2e** and **Extended Data Fig. 6d**).

Overall, NanoSeq data on granulocytes and colonic epithelium yielded similar mutation burdens and signatures to their corresponding stem cells. While larger studies will be needed to identify subtler differences 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.

Mutagenesis in neurons and smooth muscle

Cortical neurons are a prime example of a post-mitotic tissue. This makes them a key cell type to study somatic mutagenesis in the absence of cell division, but also inaccessible to traditional sequencing methods. Despite the technical challenges impeding progress, somatic mutations in neurodegeneration have attracted considerable interest[1,12,13,29].

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 linear accumulation of 17.1 substitutions (linear regression, CI95%:13.7-20.5) and 2.5 indels (CI95%:1.7-3.3) per year, approximately constant throughout life (**Fig. 3a**, **b**, **Supplementary Table 8**). 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].

A previous study using SNP-phased error-corrected single-cell sequencing reported three signatures in neurons, one that increased linearly with age and two that did not[13]. The spectrum found by NanoSeq and the mutation rate per year closely resemble the age-associated signature in that study (cosine similarity 0.96; **Fig. 3a**, **c** and **Extended Data Fig. 7b**, **c**). The two other signatures, responsible for around 72% of all mutations reported in the study (**Extended Data Fig. 7d**), appear exclusively in single-cell data and likely derive from amplification errors or transient DNA damage. Consistent with this possibility, the dominant signature in single-neuron data closely resembles a single-cell-specific signature reported in vitro[16] (cosine similarity 0.97, **Extended Data Fig. 7b**).

To better understand the mutational processes active in post-mitotic neurons, we performed signature decomposition on NanoSeq data from neurons, 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 mutational spectra from neurons (**Fig. 3c**, **d**) showed several interesting features. T>C substitutions at ApT sites appear enriched in neurons and show strong transcriptional strand biases (**Extended Data Fig. 8b**, **c**). Signature B (SBS1), which is believed to be caused by 5-methylcytosine deamination and fixed during DNA replication, accumulates at a low rate with age in neurons (2.5 substitutions per year, linear regression CI95% 0.9-4.1, $P = 0.005$; **Extended Data Figs. 7e** and **9a**, **b**). This suggests that 5-methylcytosine deamination can be fixed in both DNA strands without cell division, possibly by DNA repair. Neurons also have a higher proportion of indels than other tissues, with an unusual enrichment of indels longer than 1bp in highly-expressed genes, a pattern that resembles a mutational process recently described in cancer genomes[31] (**Fig. 3d**, **f** and **Extended Data Fig. 9c**, **d**). In contrast to other somatic tissues, neurons did not exhibit a clear association between expression levels and substitution rates across genes (**Fig. 3g**) and the enrichment of mutations in heterochromatin was weaker (**Fig. 3h** and **Extended Data Fig. 8a**).

Although the difference is small, AD donors showed a slightly lower substitution rate than healthy donors (linear regression, 17.6 (CI95%:15.0-20.2) vs 19.9 (CI95% 16.8-23.0) substitutions/year, $P = 0.0029$) (**Fig. 3i**, **Extended Data Fig. 7e**, **Supplementary Table 8**). This could simply 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]. We used laser microdissection of histological sections of bladder and colon to collect smooth muscle from 10 donors (**Supplementary Table 1**, **2**, **Extended Data Figs. 6b** and **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 ~20.7 substitutions per year per diploid genome (CI95%:13.7-28.0) and ~1.3 indels per year (95%:0.4-2.3) (**Fig. 3j,k**, **Supplementary Table 8**). Despite their different anatomical origin, smooth muscle cells from the bladder and colon walls showed relatively similar mutation rates. Overall, the mutational spectrum of smooth muscle shared some similarities with that of granulocytes and neurons (**Figs. 3l-n** and **1c**), with all three signatures (A-C) accumulating linearly with age (**Extended Data Fig. 7f**). The 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 variation in endogenous mutagenesis across individuals is modest, at least in the cohorts studied here.

Discussion

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

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 study the mutational landscape of non-dividing cells in mitotically-active and inactive tissues. 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**, **Supplementary Note 9**). 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, with similar rates and signatures to some mitotically-active tissues, indicates 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, division-independent mutational processes may play a larger role in adult 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 in polyclonal samples, across tissues and conditions. Being insensitive to clonality, NanoSeq can also be used to efficiently and accurately quantify somatic mutation rates and signatures in 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 Xing, D., Tan, L., Chang, C. H., Li, H. & Xie, X. S. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2013106118 (2021).

16 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).

17 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).

18 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).

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

20 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).

21 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).

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 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).
27 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).
28 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).
29 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).
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).

Figure legends

Figure 1 | **BotSeqS and NanoSeq sequencing protocols**. **a**, Duplex sequencing protocol. **b**, BotSeqS mutation burden estimates in cord blood granulocytes compared to 100 single-cell derived blood colonies from two donors. **c**, BotSeqS and NanoSeq substitution profiles for cord blood granulocytes, and cosine similarities (**Methods**) with the cord blood colonies profile. **d**, Substitution imbalances are present in standard BotSeqS protocols but absent from NanoSeq (**Extended Data Figs. 1** and **2** show further details for a library of granulocytes from a 59-year-old donor). Imbalances were tested with a binomial test (*p* of 0.5) and p-

values were corrected with Benjamini and Hochberg's FDR method. **e**, Standard BotSeqS (top) and NanoSeq protocols (bottom) for library preparation. **f**, **g**, NanoSeq mutation burden estimates for cord blood granulocytes (S1/PD48442, n=6 libraries; S2/PD47269, n=1) and sperm from a 21-year-old donor (n=7) compared to blood colonies and sperm estimates, respectively; **b**, **f**, **g**, Bars show point estimates and their 95% Poisson confidence intervals. **b**, **f**, Box plot shows the interquartile range, median, 95% confidence interval for the median, and outliers as grey dots. **f**, The mean and its 95% confidence interval are shown in red.

Figure 2 | Mutation in stem and 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 granulocytes and single-cell derived blood colonies. NanoSeq estimates for granuloctyes (red dots) obtained for one library per donor except for donors of ages 54 (n=2), 63 (n=2), and 59 (n=5). Standard sequencing estimates are shown as box plots and based on 10 colonies per donor, except for the 59-year-old (n=110) and cord blood (n=100). **c**, Granulocytes and blood colonies substitution profiles and their cosine similarity (**Methods**) for the 59-year-old donor. **d**, Burden estimates in colonic crypts from three donors, comparing standard methods (box plots) and NanoSeq (red dots; n=3, 2 and 2 libraries per donor); **e**, Accumulation of substitutions throughout life in colonic crypts from five donors, excluding substitutions attributed to the episodic colibactin signature. **b**, **d**, Dots and lines show point estimates and their corresponding 95% Poisson confidence intervals, respectively. **b**, **d**, Box plots show the interquartile range, median, 95% confidence interval for the median, with outliers as grey dots. **b**, **e**, Linear mixed regression models for granulocytes (red dashed line), blood colonies (dark cyan), and colonic crypts (black), with 95% confidence intervals calculated through parametric bootstrapping (**Methods**). Regression intercepts and slopes are provided in **Supplementary Table 8**.

Figure 3 | Mutation landscape in neurons and smooth muscle. **a**, **b**, Accumulation of substitutions and indels in neurons throughout life, including healthy (n=8) and Alzheimer's disease (n=9) donors. **c**, **d**, Substitution and indel spectra in neurons; a description of each type of indel can be found in **Extended Data Fig. 5d**. **e**, Signature decomposition. **f**, **g**, Indel and substitution rates in genes in the whole cohort by level of expression. **h**, Substitution rates in transcribed and quiescent/heterochromatin DNA across different cell types (spectra in **Extended Data Fig. 8a**). **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 (n=2 libraries for donors aged 54 and 68). **l**, **m**, Substitution and indel spectra in smooth muscle. **n**, Exposure to signatures A, B and C in smooth muscle samples. **o**, **p**, Substitution and indel accumulation per year across different cell types; 95% confidence intervals estimated through simple (neurons and urothelium) or mixed effect (rest) linear regression with intercept=0; vertical lines show regression 95% confidence intervals. **a**, **b**, **f**, **g**, **h**, **j**, **k**, vertical lines show Poisson 95% confidence intervals. **a**, **b**, Linear regression model as dashed lines, showing 95% confidence interval as grey areas. **j**, **k**, Linear mixed effect regressions as dashed lines, showing 95% confidence intervals obtained through parametric bootstrapping (**Methods**) as grey areas. Regression results with free or zero intercept are provided in **Supplementary Table 8**.

Methods

Sample collection and ethics

All samples were collected with informed consent from all human research participants or their families. The haematological samples in the study were obtained from 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 were collected under REC ethics approval EC04/015, London - Westminster REC; 16/NE/003, NRES Committee North East-Newcastle and North Tyneside 1. Colon and bladder tissue were collected by the Cambridge Biorepository for Translational Medicine (REC reference: 15/EE/0152 NRES Committee East of England – Cambridge South). Frozen biopsies of frontal cortex from healthy and 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).

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[26] 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 were 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[26]. 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*.[26]. For the remaining donors, whole blood was diluted with PBS and mononuclear cells (MNC) were isolated using lymphoprep™ (STEMCELL Technologies) density gradient centrifugation. The MNC fraction was then removed to a fresh tube, leaving behind the red cell pellet, which also contained the granulocyte fraction. The MNC fraction was depleted of red blood cells by a single 15 min incubation with RBC lysis buffer (BioLegend) at 4^{o}C. Granulocytes were purified from the red cell pellets using 3 incubations (for 20 mins/10 mins/10 mins respectively) with RBC lysis buffer (BioLegend) at room temperature. 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 (1:500), CD90/PE (1:50), CD49f/PECy5 (1:100), CD38/PECy7 (1:100), CD19/A700 (1:300), CD34/APC Cy7 (1:100), CD45RA/BV421 (1:100), and Zombie/Aqua (1:2000).

Cells were stained (30 minutes at 4°C) 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).

Sperm samples

DNA was extracted from sperm samples from two donors, aged 21 and 73 years, and sequenced using the NanoSeq protocol. 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. 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 Figs. 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

Neuronal nuclei were isolated, stained and extracted from the frontal cortex samples as per Krishnaswami et al.[40] using frozen biopsies of frontal cortex from eight healthy and nine Alzheimer's disease donors. 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 (1:500; 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 shown in Fig. 1 and Extended Data Fig. 3 were prepared as follows: DNA was sheared to 450 bp using a Covaris. DNA was cleaned up using a 2.5X Ampure XP (Beckman Coulter) bead ratio. DNA was eluted in 12 µL NFW. 10 µL of the elution product were taken into the ligation reaction consisting in addition of 3.74 µL NEBuffer 4, 3.74 µL 10 mM ATP, 0.33 µL xGen Duplex Seq Adapters (IDT 1080799), 0.56 µL T4 DNA ligase (NEB M0202L) and 19.03 µL NFW. The reaction was incubated at 20 ºC for 20 min. The DNA was cleaned-up using 37.4 µL Ampure XP beads and DNA was eluted in 50 µL NFW. Libraries were quantified (qPCR) and amplified following the NanoSeq protocol. For the BotSeqS data on granulocytes from a 59-year-old donor, shown in Extended Data Figs. 1 and 2, we used an earlier implementation of the protocol. 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 NFW. 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 equimolar dATP/ddBTPs (**Supplementary Note 3**), and 1.85 µL NFW. 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 NFW. 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.

Mung Mean NanoSeq libraries were prepared as follows: DNA was sheared to an average size of 450 bp using focused ultrasonication (Covaris 644 LE220). Sheared DNA was quantified and 50 ng were used as input per reaction. Mung Bean nuclease (NEB: M0250S) was diluted to 1U, 0.5U or 0.25U/ µL in 1X Mung Bean nuclease buffer. The Mung Bean reaction was carried out in a final volume of 30 µL including 2.9 µL 10X Mung Bean nuclease buffer, 1 µL diluted Mung Bean nuclease, 10 µL DNA and 16.1 µL NFW. The reaction was incubated at 30 ºC for 30 min. Then, 1 µL 0.3% SDS was added and the reaction was cleaned up using 77.5 µL Ampure XP beads. Samples were eluted in 12 µL NFW. 10 µL was used as input into a phosphorylation reaction by adding 1.5 µL NEBuffer 4 (NEB B7004S), 1.5 µL 10 mM ATP (Fisher Scientific 10304340), 0.6 µL T4 Polynucleotide Kinase (NEB M0201S) and 1.4 µL NFW. The reaction was incubated at 37 ºC for 30 min. 13 µL were taken forward into an A-tailing reaction, adding 0.2 µL NEBuffer 4, 1.5 µL 1 mM dATP/ddBTP (NEB N0440S/GE Healthcare 27204501), 0.15 µL Klenow fragment (3'→5' exo-, NEB M0212L) and 0.15 µL NFW. The reaction was incubated at 37 ºC for 30 min. The whole 15 µL were taken into the ligation reaction mix, which consisted of 2.24 µL NEBuffer 4, 3.74 µL 10 mM ATP, 0.33 µL xGen Duplex Seq Adapters (IDT 1080799), 0.56 µL T4 DNA ligase (NEB M0202L) and 15.53 µL NFW. The reaction was incubated at 20 ºC for 20 min. The DNA was cleaned-up using 37.4 µL Ampure XP beads and DNA was eluted in 50 µL NFW. Libraries were quantified and amplified following the NanoSeq protocol.

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 (~300 x 10^6 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, pre-processing 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, on 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 $j = 1...d$:

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

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:

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

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 same filters were applied to NanoSeq and BotSeq data. (a) 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. (b) 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). (c) 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 (for sites where the two mates overlap the minimum of the average AS-XS for forward and reverse mates is taken). This filter is essential to remove mapping artefacts and a minimum AS-XS of 50 is applied also to the matched normal. (d) The average number of mismatches (NM) in a group of reads (forward or reverse) should not be higher than 2, either in the matched normal or the sample at hand. To avoid a bias in the filtering of mutation and reference calls, where a consensus base call is different from the reference, mismatches from that call are not considered when calculating the number of mismatches in the read. For sites where the two mates overlap, the maximum of the average NM for forward and reverse mates is taken. (e) No 5′ clips are allowed. (f) No improper pairs are allowed in the read bundle to avoid unreliable mappings. (g) 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. (h) Reads in the read bundle must contain no indels (except for indel calling). (i) 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 five reads aligned to each strand. (k) 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. (l) 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.

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: 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 the following three bcftools commands: 1) bcftools index -f genotype_likelihods.bcf genotype_likelihods.indexed.bcf; 2) bcftools call --skip-variants snps --multiallelic-caller --variants-only -O v genotype_likelihods.bcf -o bcftools.tmp.vcf; and 3) 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 the mask, 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, including neurons, colonic crypts and smooth muscle samples. 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 potential contaminants and that the maximum support from any one matched normal was lower than 3 reads. All the samples from plates showing evidence of contamination are considered as potential contaminants. Thresholds to remove contaminant calls 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.

Indels were not analysed for nine samples with signs of contamination as we did not implement a decontamination pipeline for indels (**Supplementary Table 4**).

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[26]. 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. (a) A custom filter was used to remove artefacts associated with the 'low input' library preparation used, including those due to cruciform DNA structures. (b) A binomial filtering strategy was used to remove variants with aggregated count distributions consistent with germline single nucleotide polymorphisms. (c) 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. (d) Sites with a mean depth below 8 and over 40 were removed. (e) 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. (f) 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**, **Supplementary Note 7**).

Bootstrapped cosine similarity

Cosine similarities are frequently used to compare mutational profiles, although they do not consider 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.

For neurons and urothelium, with only one sample per donor, we used simple multiple linear regressions (**Supplementary Table 8**), while for the remaining cell types with multiple samples per donor (smooth muscle, colonic crypts, blood colonies, granulocytes and sperm) we used linear mixed-effect models, using donor as a random effect.

For simplicity, in the comparison of substitution and indel rates per year across all cell types shown in **Fig. 3o**,**p**, we used regression models without a free intercept, after verifying that the estimated intercepts were not significantly different from zero. All the regression models, with and without intercepts, and their parameter estimates are summarised in **Supplementary Table 8**.

To test for the significance of a given fixed effect (such as organ of origin), we used the anova R function, comparing the null model without the fixed effect and the alternative model with the fixed effect (**Supplementary Table 8**). Confidence intervals for linear mixed-effects models at different ages 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.

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. Sperm samples are available under EGAC0000100027. Standard sequencing data has been deposited in EGA under accession number EGAD00001006595. For samples publicly available, references to the original sources are provided in **Supplementary Table 1**. Substitution and indel rates are available in **Supplementary Table 4**. Substitution and indel calls for samples sequenced with NanoSeq are available in **Supplementary Tables 5** and **6**. Trinucleotide substitution profiles are available in **Supplementary Table 7**.

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, perform signature extraction and signature fitting with sigfit, simulate the efficiency of the NanoSeq protocol, calculate mutation burden in specific genomic regions, and to reproduce most of the main plots are available from https://github.com/fa8sanger/NanoSeq_Paper_Code. Analyses in R were done with R v3.3 and v3.6. R libraries used include: GenomicRanges[53] (v1.38.0), Rsamtools (v2.2.3), MASS (v7.3-51.5), sigfit[52] (v2.0), readxl (v1.3.1), deconstructSigs (v1.8.0), lsa (v0.73.2), deepSNV[54] (v1.32.0), lme4 (v1.1-26), afex (v0.28-1), lmerTest (v3.1-3), bootpredictlme4 (v0.1), and Biostrings (v2.54.0). Our pipeline makes use of samtools[55] v1.9, bcftools[56] v1.9, bwa v0.7.5a-r405, and bedtools[57] v2.29.0. We also used the following programs: CaVeMan (v 2020), Pindel (v 2020), and MPBoot 1.1.0.

Method 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 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).
54 Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198-1204, doi:10.1093/bioinformatics/btt750 (2014).
55 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
56 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
57 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
58 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

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. We thank the participants and local coordinators at the TwinsUK.

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). The TwinsUK study was funded by the Wellcome Trust and European Community's Seventh Framework Programme (FP7/2007-2013). The TwinsUK study also receives support from the National Institute for Health Research (NIHR)- funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

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.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 has been filed including R.J.O., F.A. and I.M.

Additional Information

Supplementary information. The online version contains supplementary material available at xxx

Correspondence and requests for materials should be addressed to R.J.O. or I.M.

Extended data figures legends

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. 2c** 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 | Mung Bean NanoSeq. **a**, 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 two protocols that are equally affected by this limitation. **b**, Substitution profiles for the standard end repair protocol (BotSeqS) and for Mung Bean, showing the cosine similarities with the reference profile (**Fig. 1c**).

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; box plots show the interquartile range, median and 95% confidence interval for the median. **c**, Comparison of indel rates between cord blood colonies (indels were called with the Pindel algorithm) and granulocytes from neonates (NanoSeq pipeline), showing Poisson 95% confidence intervals. Given the sparsity of indel calls in cord blood, data from different colonies (n=100) and granulocytes (n=2 donors, one of them with 5 replicates) were combined into single point estimates. **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. For donor PD34200 a single crypt, a pool of six crypts, and two smooth muscle areas were sequenced. For donor PD37449, the two single crypts and the pool of six crypts were sequenced. The burden estimates for these microbiopsies are shown in **Fig. 2c** and **3j**, **k**.

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.*[16]). **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[58], using the following epigenomes: E073 (frontal cortex), E030 (granulocytes), E076 (smooth muscle), and E075

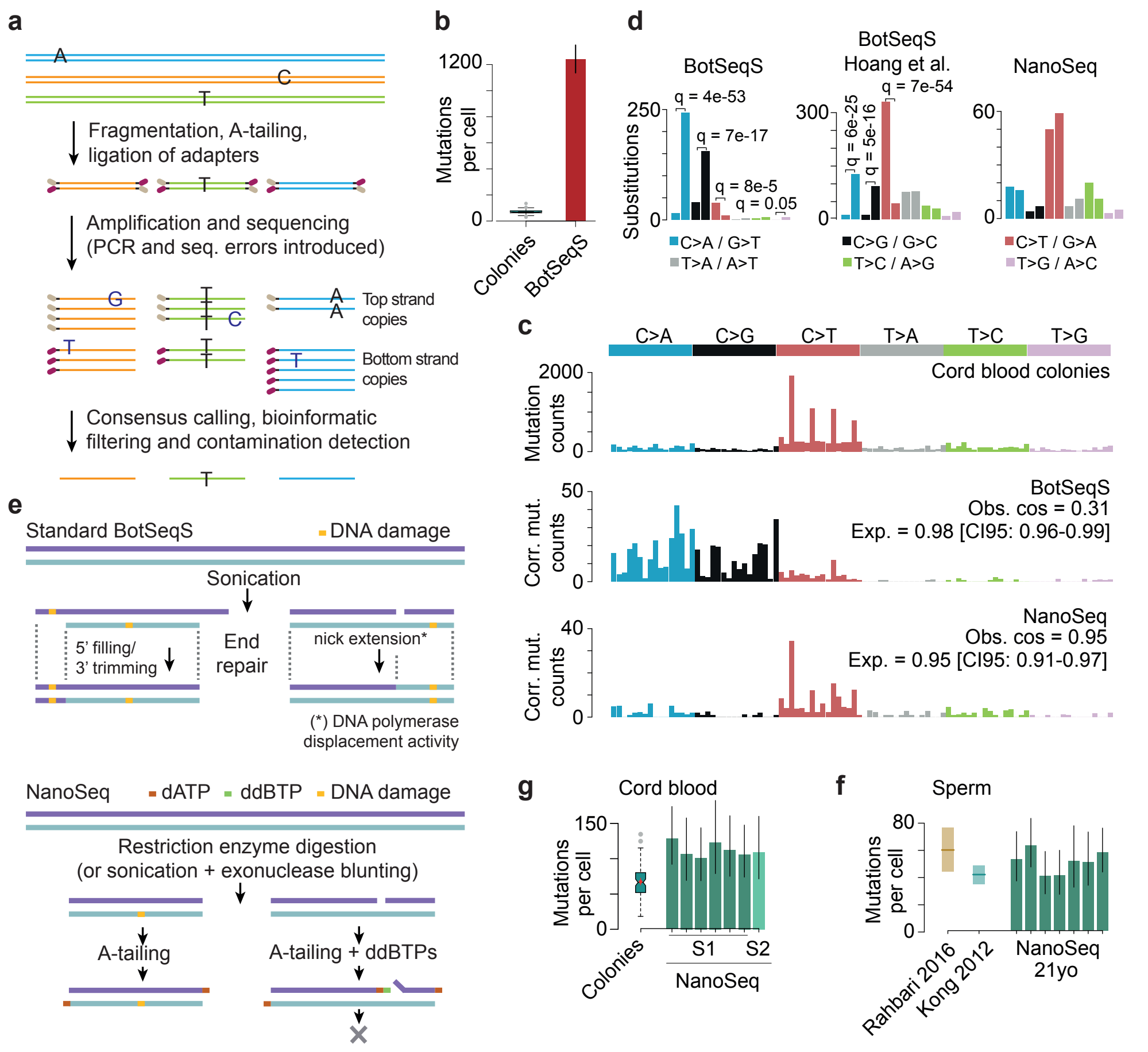
(colonic mucosa). To enable direct comparison of spectra across genomic regions with different trinucleotide frequencies, the profiles have been normalised to the genomic trinucleotide frequencies (**Methods**). **b**, Transcriptional strand asymmetries in neurons, granulocytes, smooth muscle and colonic crypts. **c**, Transcriptional strand asymmetries in neurons in quartiles of gene expression.

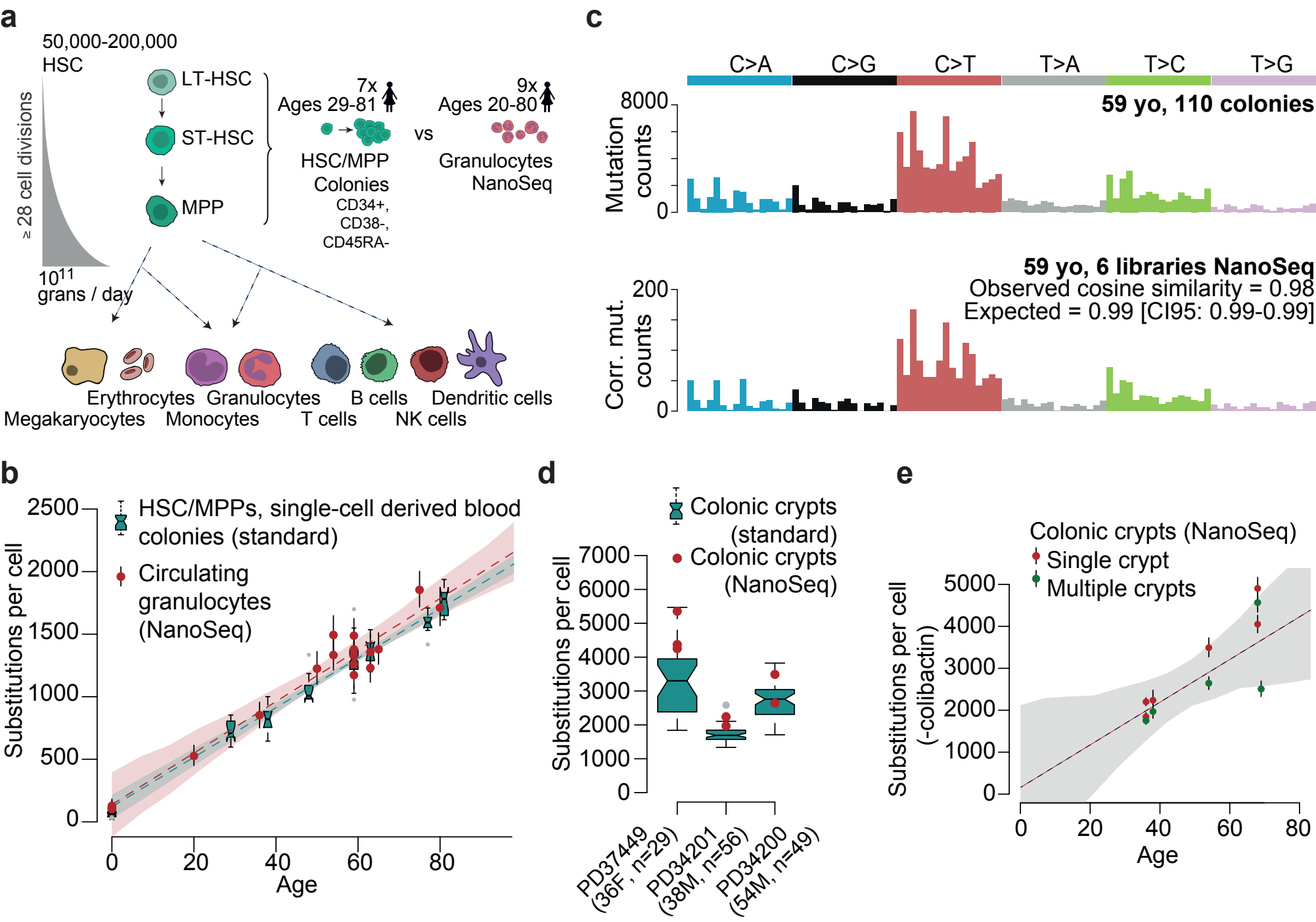
Extended Data Figure 9 | Additional substitution and indel spectra. **a**, NanoSeq mutational spectrum for neurons corrected for trinucleotide frequency in the callable genome. Unlike the usual representation, which shows unnormalized rates, this representation shows mutation rates per available trinucleotide. **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].

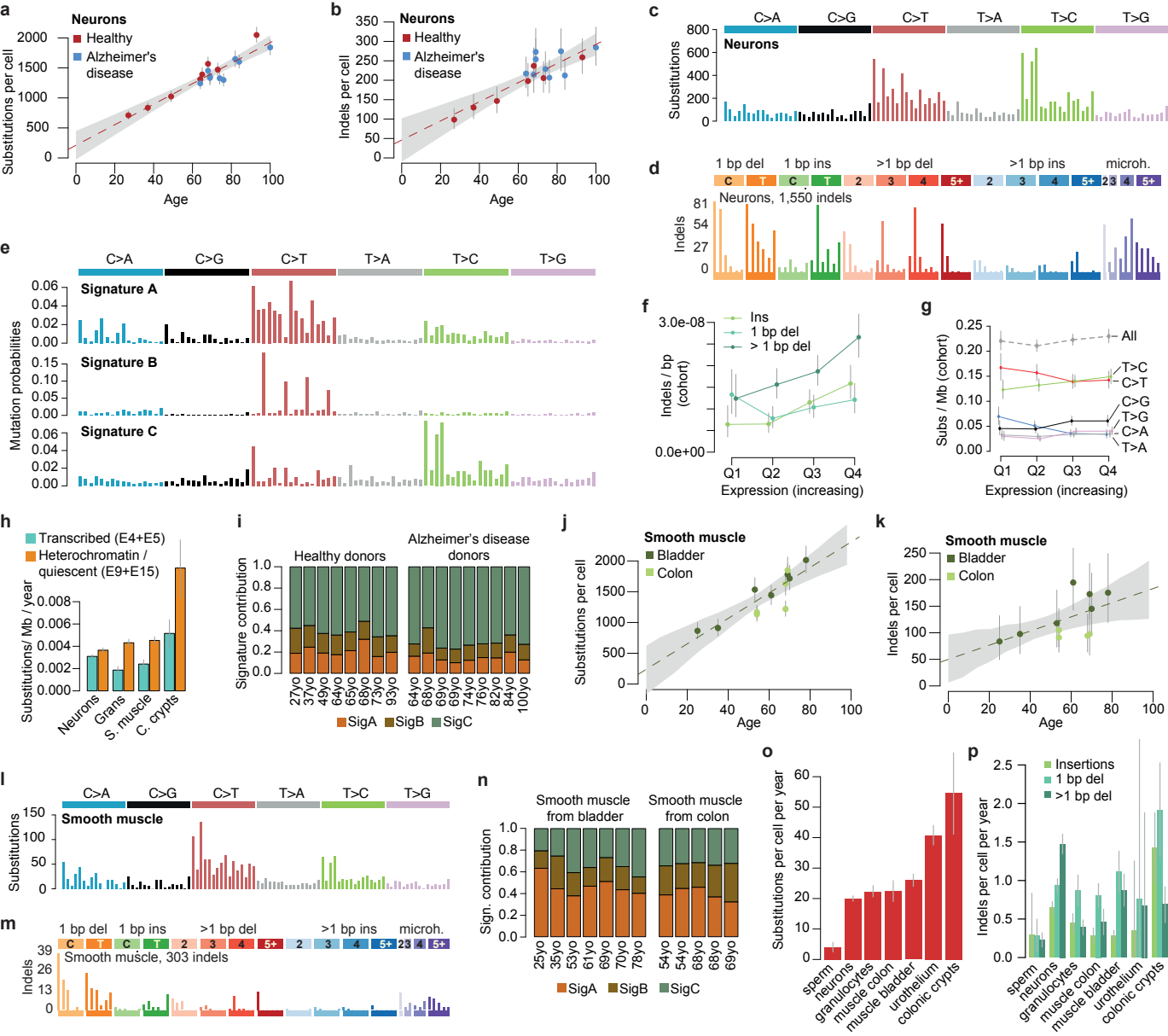
Extended Data Figure 10 | Smooth muscle. **a**, Histology of bladder smooth muscle showing two sections from donor PD40842; only one of the two sections was sequenced with NanoSeq. **b**, Number of mutations detected with CaVEMan in different smooth muscle sections processed with our standard microdissection 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; box plots show the interquartile range, median, 95% confidence interval for the median, and outliers as black dots. Box plot notches show the 95% confidence interval for the median.

Supplementary tables legends

Supplementary Table 1. Samples used in this study and corresponding data availability
Supplementary Table 2. Sequencing yields for NanoSeq/BotSeqS DNA libraries
Supplementary Table 3. *In silico* restriction enzyme digestion of the human genome
Supplementary Table 4. Substitution and indel rates
Supplementary Table 5. Substitution calls (NanoSeq protocol)
Supplementary Table 6. Indel calls (NanoSeq protocol)
Supplementary Table 7. Trinucleotide substitution profiles
Supplementary Table 8. Linear regression models


a
A
C
T
Fragmentation, A-tailing,
ligation of adapters
Amplification and sequencing
(PCR and seq. errors introduced)
G
C
A
A
Top strand
copies
T
T
Bottom strand
copies
Consensus calling, bioinformatic
filtering and contamination detection
b
1200
Mutations
per cell
0
Colonies
BotSeqS
d
BotSeqS
BotSeqS
Hoang et al.
NanoSeq
Substitutions
250
0
q = 4e-53
q = 7e-17
q = 8e-5
q = 0.05
300
0
q = 6e-25
q = 5e-16
q = 7e-54
60
0
C>A / G>T
T>A / A>T
C>G / G>C
T>C / A>G
C>T / G>A
T>G / A>C
c
C>A
C>G
C>T
T>A
T>C
T>G
Cord blood colonies
Mutation
counts
2000
0
BotSeqS
Obs. cos = 0.31
Exp. = 0.98 [CI95: 0.96-0.99]
Corr. mut.
counts
50
0
NanoSeq
Obs. cos = 0.95
Exp. = 0.95 [CI95: 0.91-0.97]
Corr. mut.
counts
40
0
e
Standard BotSeqS
DNA damage
Sonication
5' filling/
3' trimming
End
repair
nick extension*
(*) DNA polymerase
displacement activity
NanoSeq
dATP
ddBTP
DNA damage
Restriction enzyme digestion
(or sonication + exonuclease blunting)
A-tailing
A-tailing + ddBTPs
g
Cord blood
Mutations
per cell
150
0
Colonies
S1
S2
NanoSeq
f
Sperm
Mutations
per cell
80
0
Rahbari 2016
Kong 2012
NanoSeq
21yo


a
50,000-200,000
HSC
LT-HSC
ST-HSC
MPP
≥ 28 cell divisions
10^11
grans / day
7x
Ages 29-81
HSC/MPP
Colonies
CD34+,
CD38-,
CD45RA-
vs
9x
Ages 20-80
Granulocytes
NanoSeq
Erythrocytes
Granulocytes
B cells
Dendritic cells
Megakaryocytes
Monocytes
T cells
NK cells
c
C>A
C>G
C>T
T>A
T>C
T>G
59 yo, 110 colonies
Mutation counts
8000
0
59 yo, 6 libraries NanoSeq
Observed cosine similarity = 0.98
Expected = 0.99 [CI95: 0.99-0.99]
Corr. mut. counts
200
0
b
HSC/MPPs, single-cell derived blood colonies (standard)
Circulating granulocytes (NanoSeq)
Substitutions per cell
2500
2000
1500
1000
500
0
0
20
40
60
80
Age
d
Colonic crypts (standard)
Colonic crypts (NanoSeq)
Substitutions per cell
7000
6000
5000
4000
3000
2000
1000
0
PD37449 (36F, n=29)
PD34201 (38M, n=56)
PD34200 (54M, n=49)
e
Colonic crypts (NanoSeq)
Single crypt
Multiple crypts
Substitutions per cell (-colibactin)
5000
4000
3000
2000
1000
0
0
20
40
60
80
Age


a
Neurons
Healthy
Alzheimer's disease
Substitutions per cell
Age
b
Neurons
Healthy
Alzheimer's disease
Indels per cell
Age
c
C>A
C>G
C>T
T>A
T>C
T>G
Neurons
Substitutions
d
1 bp del
1 bp ins
>1 bp del
>1 bp ins
microh.
Neurons, 1,550 indels
Indels
e
C>A
C>G
C>T
T>A
T>C
T>G
Signature A
Signature B
Signature C
Mutation probabilities
f
Ins
1 bp del
> 1 bp del
Indels / bp (cohort)
Expression (increasing)
g
All
T>C
C>T
C>G
T>G
C>A
T>A
Subs / Mb (cohort)
Expression (increasing)
h
Transcribed (E4+E5)
Heterochromatin / quiescent (E9+E15)
Substitutions/ Mb / year
Neurons
Grans
S. muscle
C. crypts
i
Healthy donors
Alzheimer's disease donors
Signature contribution
SigA
SigB
SigC
j
Smooth muscle
Bladder
Colon
Substitutions per cell
Age
k
Smooth muscle
Bladder
Colon
Indels per cell
Age
l
C>A
C>G
C>T
T>A
T>C
T>G
Smooth muscle
Substitutions
m
1 bp del
1 bp ins
>1 bp del
>1 bp ins
microh.
Smooth muscle, 303 indels
Indels
n
Smooth muscle from bladder
Smooth muscle from colon
Sign. contribution
SigA
SigB
SigC
o
Substitutions per cell per year
sperm
neurons
granulocytes
muscle colon
muscle bladder
urothelium
colonic crypts
p
Insertions
1 bp del
>1 bp del
Indels per cell per year
sperm
neurons
granulocytes
muscle colon
muscle bladder
urothelium
colonic crypts