Article

DNA mismatch and damage patterns revealed by single-molecule sequencing

https://doi.org/10.1038/s41586-024-07532-8

Received: 19 February 2023

Accepted: 7 May 2024

Published online: 12 June 2024

Check for updates

Mei Hong Liu[1,2,10], Benjamin M. Costa[1,2,10], Emilia C. Bianchini[1,2], Una Choi[1,2], Rachel C. Bandler[1], Emilie Lassen[3], Marta Grońska-Pęski[1,2], Adam Schwing[1,2], Zachary R. Murphy[1,2], Daniel Rosenkjær[3], Shany Picciotto[4], Vanessa Bianchi[5], Lucie Stengs[5], Melissa Edwards[5], Nuno Miguel Nunes[5], Caitlin A. Loh[1,2], Tina K. Truong[1,2], Randall E. Brand[6], Tomi Pastinen[7], J. Richard Wagner[8], Anne-Bine Skytte[3], Uri Tabori[5,9], Jonathan E. Shoag[4] & Gilad D. Evrony[1,2]✉

Mutations accumulate in the genome of every cell of the body throughout life, causing cancer and other diseases[1,2]. Most mutations begin as nucleotide mismatches or damage in one of the two strands of the DNA before becoming double-strand mutations if unrepaired or misrepaired[3,4]. However, current DNA-sequencing technologies cannot accurately resolve these initial single-strand events. Here we develop a single-molecule, long-read sequencing method (Hairpin Duplex Enhanced Fidelity sequencing (HiDEF-seq)) that achieves single-molecule fidelity for base substitutions when present in either one or both DNA strands. HiDEF-seq also detects cytosine deamination—a common type of DNA damage—with single-molecule fidelity. We profiled 134 samples from diverse tissues, including from individuals with cancer predisposition syndromes, and derive from them single-strand mismatch and damage signatures. We find correspondences between these single-strand signatures and known double-strand mutational signatures, which resolves the identity of the initiating lesions. Tumours deficient in both mismatch repair and replicative polymerase proofreading show distinct single-strand mismatch patterns compared to samples that are deficient in only polymerase proofreading. We also define a single-strand damage signature for APOBEC3A. In the mitochondrial genome, our findings support a mutagenic mechanism occurring primarily during replication. As double-strand DNA mutations are only the end point of the mutation process, our approach to detect the initiating single-strand events at single-molecule resolution will enable studies of how mutations arise in a variety of contexts, especially in cancer and ageing.

Mosaic mutations are ubiquitous in the body and accumulate throughout life in every cell[1,2]. Most mosaic mutations begin as nucleotide mismatches or damage in only one of the two strands of the DNA double helix[3,4]. When these single-strand DNA (ssDNA) events are misrepaired, or when they are replicated during the cell cycle before repair, they then become permanent double-strand DNA (dsDNA) mosaic mutations[3]. Although current methods for profiling mosaic changes to DNA achieve high fidelity for dsDNA mutations, they cannot accurately resolve these precursor ssDNA events. This is because current methods—single-cell genome sequencing[5], in vitro cloning of single cells[6], microdissection or biopsy of clonal populations[7], and duplex sequencing[8,9]—amplify the original DNA molecules before sequencing, either prior to or on the sequencer itself. This masks true ssDNA events by either transforming existing ssDNA mismatches and damage to dsDNA mutations, or by introducing artefactual ssDNA mismatches and damage[8].

Mosaic dsDNA mutations are the result of the interaction between ssDNA mismatch and damage events, DNA repair, and DNA replication[3]. Consequently, dsDNA mutational signatures (that is, the sequence contexts of mutations) may not reflect the patterns of the originating ssDNA events[4]. dsDNA mutation profiling also does not resolve on which strands the initiating ssDNA events occur. Therefore, a complete understanding of mutational processes requires profiling of ssDNA mismatches and damage[3,10]. Here, to study the ssDNA origins of mosaic mutations, we developed an approach for direct sequencing of single DNA molecules without any previous amplification that achieves, for

[1]Center for Human Genetics and Genomics, New York University Grossman School of Medicine, New York, NY, USA. [2]Department of Pediatrics, Department of Neuroscience & Physiology, Institute for Systems Genetics, Perlmutter Cancer Center, and Neuroscience Institute, New York University Grossman School of Medicine, New York, NY, USA. [3]Cryos International Sperm and Egg Bank, Aarhus, Denmark. [4]Department of Urology, University Hospitals Cleveland Medical Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA. [5]Program in Genetics and Genome Biology, Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, Toronto, Ontario, Canada. [6]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [7]Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA. [8]Department of Nuclear Medicine and Radiobiology, Université de Sherbrooke, Sherbrooke, Quebec, Canada. [9]Division of Haematology/Oncology, Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. [10]These authors contributed equally: Mei Hong Liu, Benjamin M. Costa. ✉e-mail: gilad.evrony@nyulangone.org

single-base substitutions, single-molecule fidelity detection of dsDNA mutations simultaneously with ssDNA mismatches and damage.

HiDEF-seq

Profiling dsDNA mosaic mutations in human tissues requires single-molecule fidelity of less than 1 error per 1 billion bases (10^{-9}), and profiling ssDNA mismatch and damage events would probably require similar or greater fidelity[8,10]. However, to our knowledge, no technology to date has achieved this fidelity when directly sequencing unamplified single DNA molecules. To achieve this, we developed HiDEF-seq. HiDEF-seq substantially increases the fidelity of single-molecule sequencing by (1) increasing the number of independent sequencing passes per strand (median of 32 passes with a median of 1.7 kilobase (kb) molecules) relative to standard single-molecule sequencing[11] to create a high-quality consensus sequence for each strand; (2) eliminating in vitro artefacts during library preparation by ssDNA nick ligation and by using either the NanoSeq A-tailing approach[8] or a protocol without A-tailing for post-mortem samples with degraded DNA; and (3) a computational pipeline that avoids analytic artefacts (Fig. 1a,b, Methods, Extended Data Figs. 1–5 and Supplementary Note 1). HiDEF-seq libraries are sequenced on Pacific Biosciences (PacBio) single-molecule, long-read sequencers. The computational pipeline analyses single-base substitutions, as these have an orthogonal error profile to the prevalent insertion and deletion sequencing errors of single-molecule sequencing[12], and it analyses each strand separately to distinguish between dsDNA and ssDNA events (Methods).

We profiled purified human sperm with HiDEF-seq as the most rigorous test of fidelity for detecting dsDNA mosaic mutations, as sperm have the lowest dsDNA mutation burden of any readily accessible human cell type[13]. Sperm dsDNA mutation burdens measured by HiDEF-seq were concordant with a previous study of de novo mutations[14] and with NanoSeq profiling[8] (a method for duplex sequencing of mosaic dsDNA mutations) that we performed for the same samples (Fig. 1c). HiDEF-seq also measured the expected dsDNA mutational signatures and linear increase in dsDNA mutation burdens with age in other human tissues (liver, kidney, blood and cerebral cortex neurons)[8,15], with one outlier blood sample of an individual with a kidney transplant (Fig. 1d, Extended Data Fig. 5i and Supplementary Note 2).

Notably, relaxing from a threshold of ≥20 to ≥5 sequencing passes per strand, while keeping our optimized computational filters, produced concordant dsDNA mutation burdens (Extended Data Fig. 3e). This suggests that PacBio sequencing can achieve a higher per-pass fidelity for substitutions than estimated by previous studies[11]. Using the probability of complementary single-strand calls occurring at the same position (Methods), we estimate HiDEF-seq's fidelity for dsDNA mutations as less than 1 error per 3×10^{13} base pairs (bp) with ≥5 passes per strand and less than 1 error per 1×10^{14} bp with ≥20 passes per strand. Accordingly, for analysis of dsDNA mutations, we used the lower threshold of ≥5 passes per strand as this increases the percentage of analysed molecules from 70% to 99.8% (of molecules passing primary data processing), and it increases the percentage of interrogated bases by 11%. HiDEF-seq uses restriction enzyme fragmentation that captures approximately 40% of the human genome (Extended Data Fig. 1a), which is sufficient for obtaining accurate mosaic mutation burdens and mutational patterns[8]. It can also use random fragmentation to enable profiling of any genomic region, although this requires more input DNA (Methods). We also successfully quantified dsDNA mutation burdens in sperm using HiDEF-seq with larger DNA fragments (median, 4.2 kb), which have correspondingly fewer (median, 15) passes per strand (Supplementary Note 3). However, for this study, we proceeded with HiDEF-seq with the smaller median 1.7 kb fragments, as a higher threshold of ≥20 passes per strand was required for ssDNA analysis.

We next analysed ssDNA calls. Importantly, these may include not only ssDNA mismatches, but also damaged bases that alter base pairing and lead to misincorporation of nucleotides by the sequencer polymerase. The latter may be advantageous as it would enable high-fidelity detection of ssDNA damage. In contrast to dsDNA mutation analysis, duplex error correction is not possible for ssDNA calls, and true ssDNA call burdens (calls per base) are unknown. Thus, for ssDNA calling, we optimized key analytic parameters by identifying filter thresholds above which ssDNA burden estimates are stable (Methods and Extended Data Fig. 3i,j). To compare ssDNA calls between HiDEF-seq and NanoSeq, we profiled 9 samples using both methods. Although HiDEF-seq and NanoSeq dsDNA mutation burdens and patterns were concordant, HiDEF-seq measured on average 18-fold lower ssDNA call burdens, with distinct patterns, and 5-fold lower when considering only C>T calls (Fig. 1e–g and Extended Data Fig. 6a–c). This suggests that, while NanoSeq achieves high fidelity for dsDNA mutations, its ssDNA calls are largely artefactual as suggested by its developers[8]. HiDEF-seq ssDNA burdens in cerebral cortex neurons were also around 13-fold lower than estimated by Meta-CS single-cell duplex sequencing[16], with a distinct pattern, and about 4-fold lower when considering only C>T calls (Supplementary Tables 2 and 3). Overall, by direct interrogation of unamplified single molecules, HiDEF-seq achieves, to our knowledge, the highest fidelity for single-base changes of any DNA-sequencing method to date.

Cancer predisposition syndromes

As there is no previous method for sequencing ssDNA mismatches with single-molecule fidelity, we sought to confirm the veracity of HiDEF-seq's ssDNA calls by profiling samples from individuals with inherited cancer predisposition syndromes that may have elevated ssDNA call burdens. We profiled 17 blood, primary fibroblast, and lymphoblastoid cell line samples from 8 different cancer predisposition syndromes, including defects in nucleotide excision repair, mismatch repair, polymerase proofreading, and base excision repair (Supplementary Tables 1 and 2). In these samples, we first confirmed HiDEF-seq's single-molecule fidelity for dsDNA mutations by measuring the expected dsDNA mutation burdens and signatures based on previous studies[17–21] (Extended Data Fig. 7a–d and Supplementary Tables 2 and 4).

Notably, compared to non-cancer predisposition samples, we detected higher ssDNA call burdens in two cancer predisposition syndromes: a 2.6-fold increase (95% confidence interval: 2.3–3.0) in *POLE* polymerase proofreading-associated polyposis syndrome samples (PPAP; germline heterozygous exonuclease domain mutations in *POLE*, which encodes the catalytic subunit of polymerase epsilon that performs leading strand genome replication[22]), and a 1.6-fold increase (95% confidence interval: 1.4–1.9) in congenital mismatch repair deficiency syndrome samples (CMMRD; *MSH2*, *MSH6*, and *PMS2* germline biallelic loss of function) (Fig. 2a). Moreover, the percentage of purine ssDNA calls (G>T/C/A and A>T/G/C) was elevated in PPAP samples (average, 61%; range, 52–73%) and CMMRD samples (average, 33%; range, 23–57%) compared to non-cancer predisposition samples (average, 20%; range, 12–29%) (Fig. 2b). In PPAP samples, this was largely due to increased G>T, G>A, and A>C ssDNA calls, while CMMRD samples exhibited smaller alterations in sequence contexts of ssDNA calls (Fig. 2b). These data indicate that most ssDNA calls in PPAP samples, and at least some calls in CMMRD samples, are bona fide ssDNA mismatches.

To further characterize the patterns of ssDNA mismatches in *POLE* PPAP samples, we plotted their 192-trinucleotide context spectra (standard 96-trinucleotide context spectra, separated by central pyrimidine versus central purine). This revealed a distinct pattern, with two large peaks for AGA>ATA and AAA>ACA accounting for around 15–20% and about 5–10% of ssDNA mismatches, respectively, in addition to smaller peaks with G>T, G>A, A>C, and C>T contexts (Fig. 2c and Supplementary Table 3). The ssDNA mismatch spectra were highly concordant with the dsDNA mutation spectra of these same samples

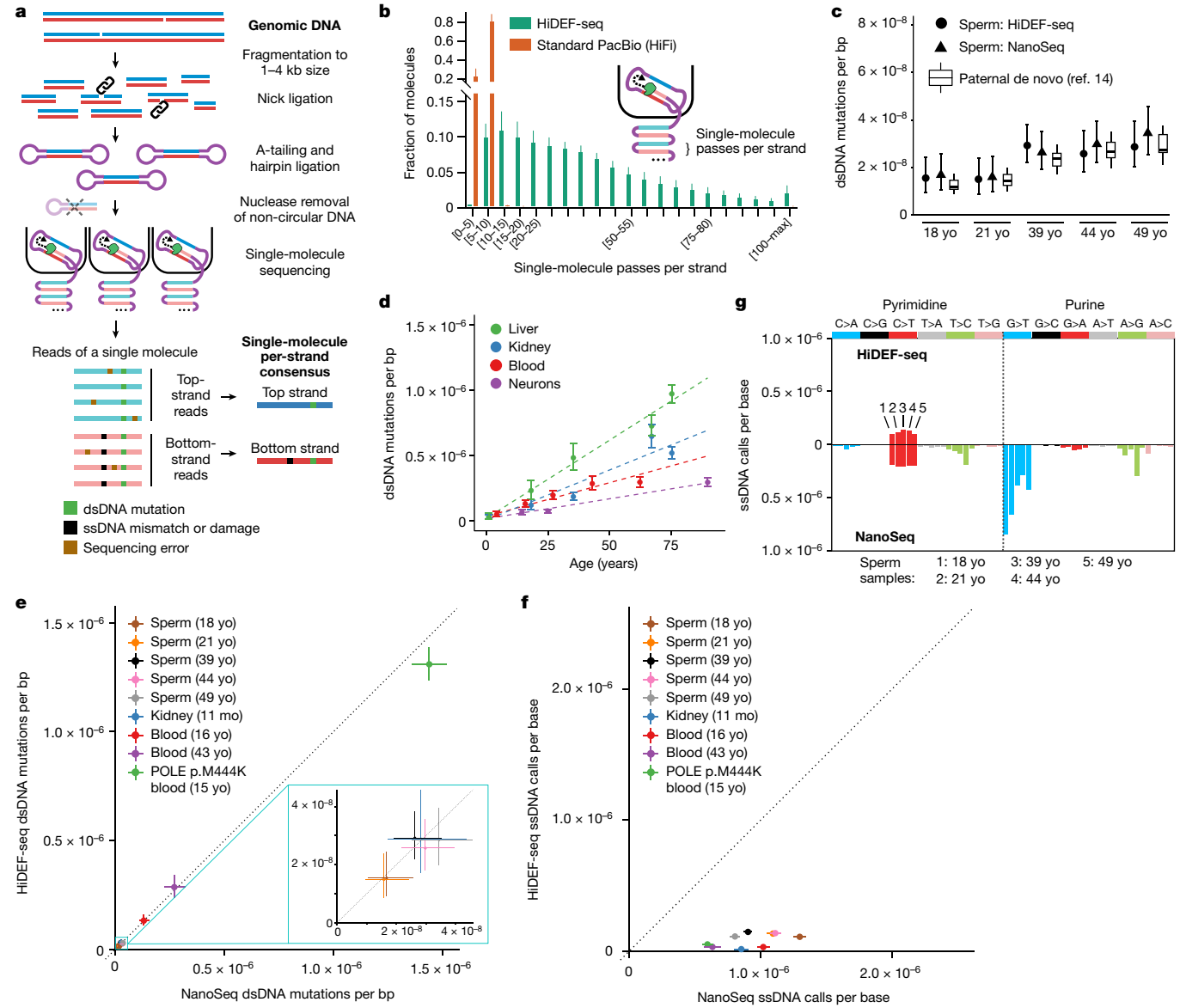


Fig. 1 | Overview of HiDEF-seq. a, HiDEF-seq schematic. A-tailing uses dATP and non-A dideoxynucleotides[8], except for lower-quality post-mortem samples that use only non-A dideoxynucleotides to avoid dATP misincorporation at residual nicks (Methods and Extended Data Fig. 5). **b**, The average fraction of molecules across representative HiDEF-seq samples ($n = 51$) and standard PacBio sequencing (HiFi) samples ($n = 10$) in different bins of number of passes per strand (Methods). The average percentage of molecules with ≥5 and ≥20 passes per strand is 99.8% and 70% for HiDEF-seq, respectively, and 78.7% and 0.1% for HiFi, respectively. The plot shows HiDEF-seq molecules output by the pipeline's primary data-processing step. *x*-axis square brackets and parentheses signify inclusion and exclusion of bin end points, respectively. **c**, HiDEF-seq and NanoSeq dsDNA mutation burdens in sperm samples (left to right, SPM-1013, SPM-1002, SPM-1004, SPM-1020, SPM-1060) were compared for each age to paternally phased de novo mutations from a previous study[14]. **d**, HiDEF-seq dsDNA mutation burdens in human tissues (Supplementary Table 1). The dashed lines show weighted least-squares linear regressions. **e**,**f**, HiDEF-seq versus NanoSeq dsDNA mutation burdens (**e**) and ssDNA call burdens (**f**). Samples are (top to bottom in legend): SPM-1013, SPM-1002, SPM-1004, SPM-1020, SPM-1060, 1443, 1105, 6501, 63143. Only sample 63143 (POLE p.M444K) is from an individual with a cancer predisposition syndrome. The dashed line shows $y = x$ (expectation for concordance). **g**, HiDEF-seq versus NanoSeq ssDNA call burdens separated by call type. For each call type, each bar represents a different sperm sample (left to right, the same samples as in **c**). **b**, Error bars show standard deviations. **c**–**f**, Dots and error bars show point estimates and their Poisson 95% confidence intervals. **c**, Box plots show the median (centre line), the first and third quartiles (box limits), and the 5% and 95% quantiles (whiskers). **c**,**e**,**f**, For each sample, HiDEF-seq and NanoSeq confidence intervals were normalized to reflect an equivalent number of interrogated base pairs (**c** and **e**) or bases (**f**) (Methods). mo, months old; yo, years old.

(Fig. 2d and Supplementary Table 4), confirming that these are true ssDNA mismatches—arising from polymerase epsilon nucleotide misincorporation—that lead to the subsequent pattern of accumulated dsDNA mutations. De novo extraction of ssDNA mismatch signatures from PPAP samples produced a signature that we name SBS10ss (SBS, single-base-substitution; ss, single-strand) (Fig. 2e). Note that we propose a nomenclature with the suffix 'ss' to distinguish between ssDNA and dsDNA signatures. Projecting SBS10ss to central pyrimidine contexts, by summing central purine and central pyrimidine spectra, produced a spectrum remarkably similar (cosine similarity = 0.97) to the dsDNA signatures extracted de novo (SBSE + SBSF) from these same samples (Fig. 2e), again indicating that the ssDNA mismatches are the inciting events leading to the dsDNA mutations. SBS10ss also had strong similarity (cosine similarity = 0.90) to COSMIC[23] SBS10c that

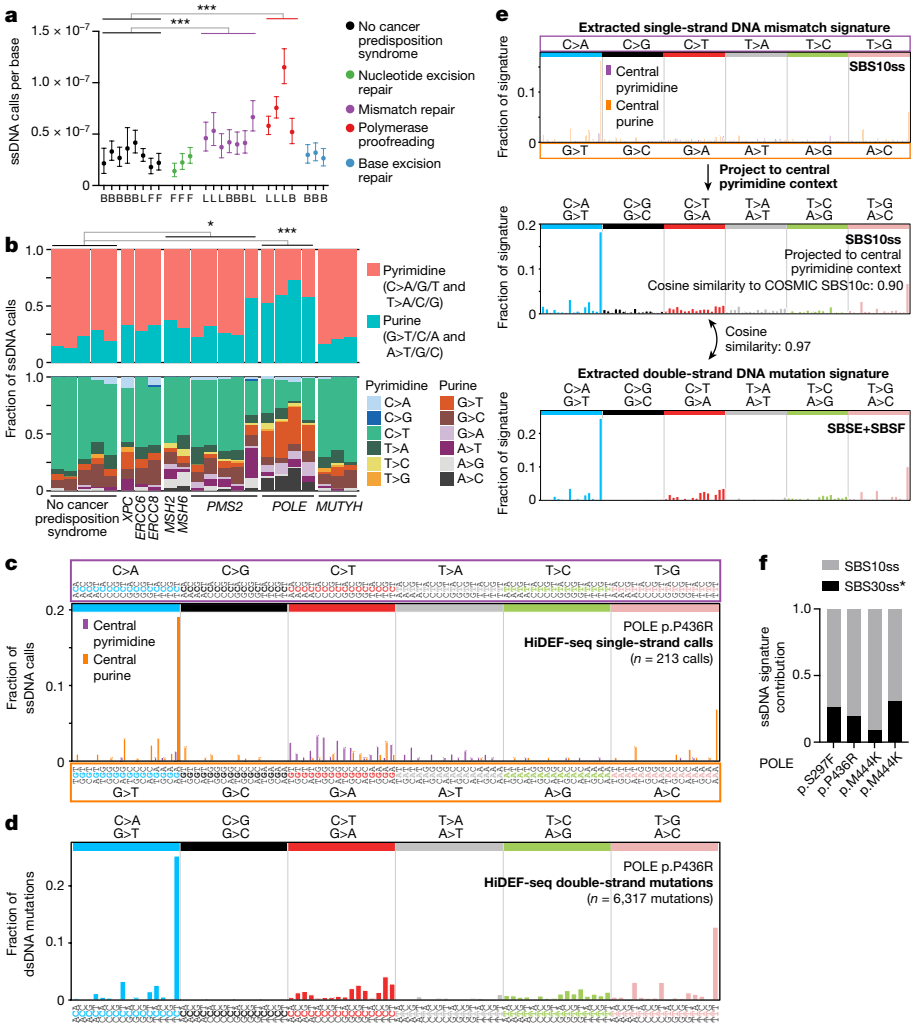


Fig. 2 | ssDNA call burdens and patterns in cancer predisposition syndromes. **a**, ssDNA call burdens in blood (B), fibroblasts (F) and lymphoblastoid cell lines (L) from individuals without and with cancer predisposition syndromes. Burdens are corrected for trinucleotide context opportunities and detection sensitivity (Methods). Statistical analysis was performed using two-sided Poisson rates ratio tests, combining calls and interrogated bases from each group, with Holm multiple-comparison adjustment; ***$P = 2 \times 10^{-10}$ for mismatch repair and $P < 10^{-15}$ for polymerase proofreading, versus non-cancer predisposition samples. Results were also significant when including only blood samples. Samples (left to right) are: 5203, 1105, 1301, 6501, 1901, GM12812, GM02036, GM03348, GM16381, GM01629, GM28257, 55838, 58801, 57627, 1400, 1324, 1325, 60603, 59637, 57615, 63143 (L), 63143 (B), CC-346-253, CC-388-290, and CC-713-555. Cancer predisposition samples are ordered as in **b**, which lists the affected genes. **b**, ssDNA call burdens by context, corrected for trinucleotide context opportunities. Statistical analysis was performed using heteroscedastic two-tailed *t*-tests, adjusted for multiple comparisons; *$P = 0.03$, ***$P = 0.0008$. Only non-cancer predisposition samples with >30 ssDNA calls were included (1105, 1301, 1901, GM12812, GM03348), as patterns are not reliably ascertained with fewer calls. However, GM16381 (*XPC*) with <30 calls was included for completeness in showing all cancer predisposition samples. **c**,**d**, Spectra of ssDNA calls (**c**) and dsDNA mutations (**d**) for representative *POLE* PPAP sample 57615, corrected for trinucleotide context opportunities. **e**, Top, the ssDNA mismatch signature SBS10ss extracted from all PPAP samples while simultaneously fitting SBS30ss* (Fig. 4d). Middle, SBS10ss projected to central pyrimidine contexts by summing central pyrimidine values and their reverse-complement central purine values to enable comparison to dsDNA signatures. Bottom, the dsDNA mutational signature (sum of SBSE and SBSF) extracted from PPAP samples. **f**, The fraction of ssDNA calls attributed to ssDNA signatures in PPAP samples (same PPAP sample order as in **a**). Cosine similarities of original spectra to spectra reconstructed from signatures (left to right) were: 0.94, 0.97, 0.97, 0.85. Sample details for **a** and **b** are provided in Supplementary Tables 1 and 2. **a**, Error bars show Poisson 95% confidence intervals.

was previously associated with *POLE* PPAP[17]. SBS10ss accounted for an average of 79% (range, 70–91%) of ssDNA calls in PPAP samples, with the remaining attributed to SBS30ss*, a ssDNA cytosine deamination damage signature (asterisk (*) indicates damage) that is described in a subsequent section (Fig. 2f). For CMMRD samples, the number of ssDNA calls was too low to extract a signature.

The two most frequent ssDNA mismatch contexts in PPAP samples are also notable for the asymmetry of their prevalence relative to their reverse complements: AGA>ATA versus TCT>TAT (73 versus 10 mismatches across all PPAP samples; χ^2 test, $P < 0.0001$) and AAA>ACA versus TTT>TGT (26 versus 2 mismatches; χ^2 test, $P < 0.0001$). These data provide a direct observation that the dsDNA mutational context AGA>ATA / TCT>TAT prevalent in *POLE* PPAP arises in vivo significantly more frequently from C:dT (template base:polymerase incorporated base) misincorporations than G:dA misincorporations, and that the dsDNA mutational context AAA>ACA / TTT>TGT arises in vivo more frequently from T:dC than A:dG misincorporations. These results are consistent with previous studies that indirectly inferred this asymmetry in yeast[24] and human tumours[25–27] harboring mutations in the polymerase epsilon exonuclease domain by identifying asymmetries

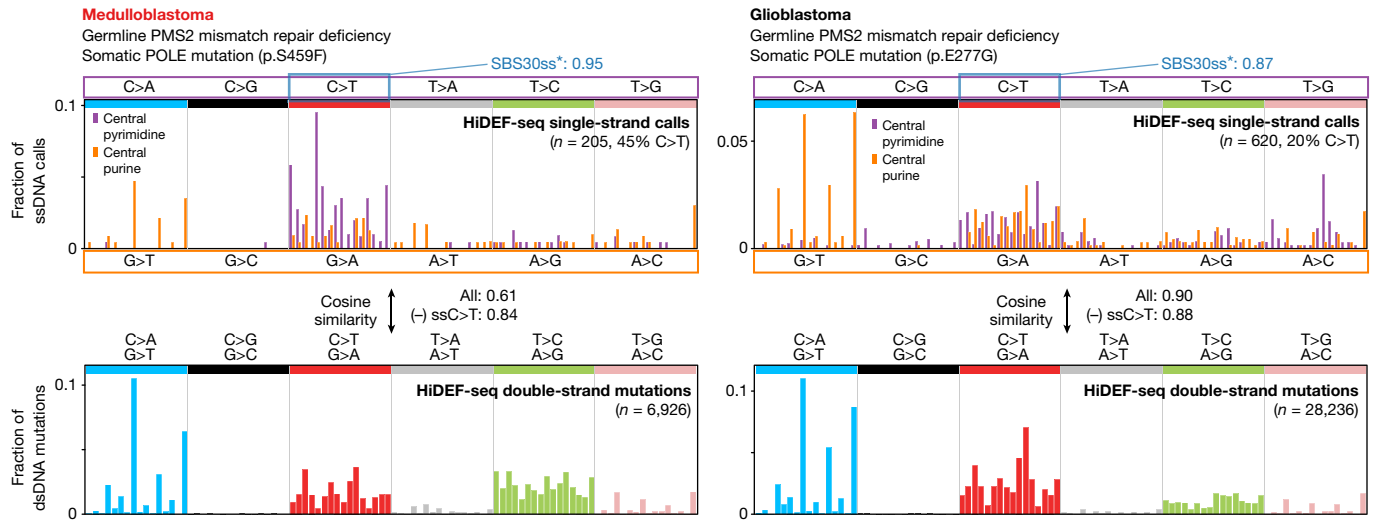


Fig. 3 | Hypermutating tumours deficient in both mismatch repair and polymerase proofreading. Spectra of ssDNA calls (top) and dsDNA mutations (bottom) in tumour samples corrected for trinucleotide context opportunities. The parentheses show the total number of raw calls and the percentage of calls that are C>T after correction for trinucleotide context opportunities. The blue annotation on the top right of each ssDNA spectrum is the cosine similarity of only the ssDNA C>T calls to SBS30ss* (details of SBS30ss* are shown in Fig. 4d). Also annotated are the cosine similarities of each sample's full ssDNA call spectrum (projected to central pyrimidine context) to its dsDNA mutation spectrum, for all ssDNA calls and after excluding ssDNA C>T calls (most of which are due to SBS30ss* cytosine deamination). Medulloblastoma ID: tumour 8; glioblastoma ID: tumour 10. Sample details are provided in Supplementary Table 1.

in the prevalence of dsDNA mutation contexts relative to their reverse complement contexts depending on whether the mutation locus is preferentially replicated through leading-strand versus lagging-strand synthesis. However, while these studies rely on replication timing data that imperfectly estimates the probability of leading- versus lagging-strand replication to measure this asymmetry, our single-molecule detection of nucleotides that were misincorporated by polymerases in vivo enables us to measure this asymmetry directly. Our results are also consistent with in vitro polymerase gap-filling assays[25,28], but, in contrast to our detection of in vivo misincorporation events, these assays lack the full context of DNA replication and repair. We also applied the above studies' indirect replication timing analysis and similarly found in our *POLE* PPAP samples a higher frequency of AGA>ATA dsDNA mutations and AGA>ATA ssDNA mismatches on the strand that is preferentially replicated in the leading direction (Extended Data Fig. 7e,f). Together, our results demonstrate direct measurements of in vivo ssDNA mismatch burdens and patterns.

Hypermutating tumours

To study the interaction between ssDNA mismatches introduced during replication and mismatch repair, we profiled three hypermutating brain tumours from individuals with CMMRD whose tumours also contained somatic mutations affecting polymerase proofreading. We excluded one tumour (tumour 3) from further analysis due to a very high ssDNA C>T burden attributed to SBS30ss* (a ssDNA cytosine deamination damage signature described in the next section) that probably arose ex vivo (Supplementary Tables 2 and 3). The other two tumours, a medulloblastoma and a glioblastoma—both with biallelic germline *PMS2* mutations and somatic *POLE* exonuclease domain mutations—had higher burdens and distinct patterns of dsDNA mutations and ssDNA calls compared with samples deficient in only mismatch repair or only polymerase proofreading (Figs. 2a–d and 3, Extended Data Figs. 7a,b and 8a–c and Supplementary Tables 2–4). Additionally, the dsDNA mutation spectra of these tumours resembled those found in previous studies of tumours and cell lines deficient in both mismatch repair and polymerase proofreading[29–32] (Fig. 3). Most dsDNA mutations were attributed to a signature with moderate similarity to COSMIC SBS14 (cosine similarity = 0.85)[31] (Extended Data Fig. 8e). Moreover, the dsDNA mutation spectra of the tumours resembled their ssDNA call spectra (Fig. 3 and Extended Data Fig. 8b,c), except for ssDNA C>T calls related to SBS30ss* (Fig. 3 and Extended Data Fig. 8f).

Importantly, the ssDNA call spectra of the tumours had notable differences relative to ssDNA call spectra of samples deficient in only polymerase proofreading, including increases in ssDNA AG>AT calls flanked by 3′ C/G/T, and increases in ssDNA G>A, A>G, and T>C calls (Figs. 2c and 3 and Supplementary Table 3). These differences in ssDNA call spectra of polymerase proofreading-deficient samples with and without mismatch repair deficiency are consistent with previous studies suggesting that mismatch repair is more efficient for certain mismatches caused by deficient polymerase proofreading[32,33]. The tumours' relative increase in ssDNA C>T calls largely arose from cytosine deamination damage rather than polymerase misincorporation (Figs. 3 and 4d and Extended Data Fig. 8f). The ssDNA call spectra further resolve the identity of the nucleotides misincorporated by proofreading-deficient polymerase epsilon—for example, C>T / G>A dsDNA mutations largely arise from C:dA rather than G:dT misincorporations (Fig. 3). We extracted a ssDNA mismatch signature from tumour samples that we name SBS14ss, as after projecting it to central pyrimidine contexts, its most similar COSMIC dsDNA signature is SBS14 (cosine similarity = 0.73 for all ssDNA calls and 0.96 for only C>A ssDNA calls) (Extended Data Fig. 8d). SBS14ss accounted for most ssDNA calls in both tumours (Extended Data Fig. 8f). We also profiled post-mortem brain and spinal cord of individuals with *MSH2* and *MSH6* CMMRD who died of brain tumours harboring somatic *POLE* mutations. This revealed not only an elevated burden of SBS1 dsDNA mutations as seen in a previous study[19], but also an elevated burden of ssDNA C>T calls at CG dinucleotides (Supplementary Note 4). This demonstrates that HiDEF-seq can also detect the ssDNA precursor lesions of SBS1 when this mutational process is elevated.

Patterns of cytosine deamination damage

A common form of DNA damage is deamination of cytosine (with or without preceding oxidation) to uracil, uracil glycol, 5-hydroxyuracil,

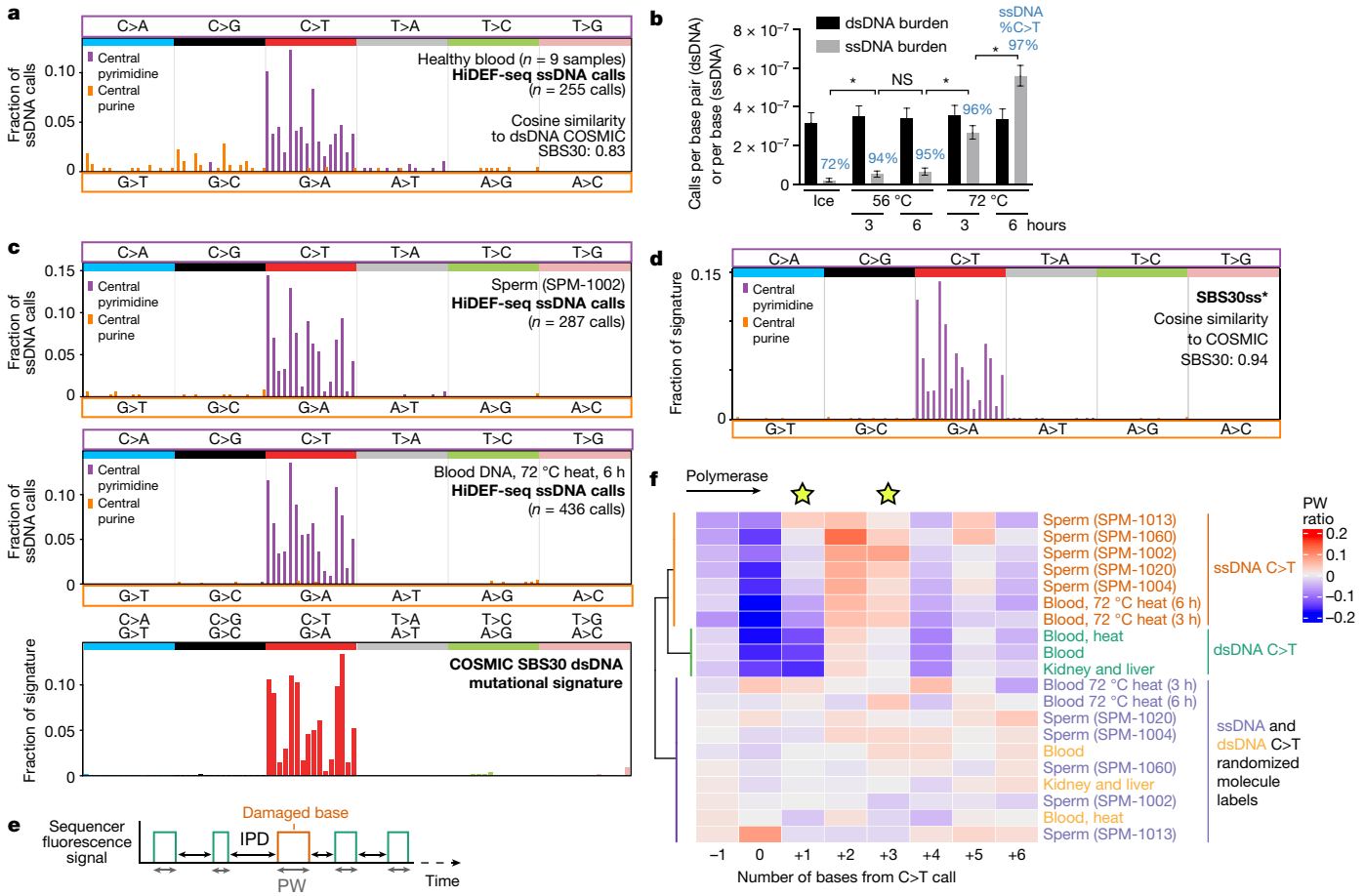


Fig. 4 | ssDNA damage signatures of sperm and heat-treated DNA. a, Spectrum of all ssDNA calls of non-cancer predisposition blood samples (one sample each from individuals 1105, 1301, 5203 and 6501, and five samples from individual 1901). The cosine similarity to COSMIC SBS30 was calculated after projecting the ssDNA spectrum to central pyrimidine contexts. **b**, dsDNA mutation and ssDNA call burdens of heat-treated DNA. **c**, ssDNA call spectra of representative sperm and heat-treated blood DNA samples, and SBS30 for comparison. **d**, SBS30ss* obtained by de novo signature extraction from central pyrimidine ssDNA calls of sperm and heat-treated samples. The cosine similarity to SBS30 was calculated after projecting to central pyrimidine contexts. **e**, Schematic of PW and IPD measured for incorporated bases during sequencing. **f**, Average PW ratios for positions −1 to +6 (relative to C>T calls), which is the polymerase footprint that has a kinetic signal that differs from the flanking baseline. Unbiased hierarchical clustering (dendrogram) separates ssDNA C>T calls from dsDNA C>T mutations and from kinetic profiles with randomized molecule labels. Positions +1 and +3 (stars) best discriminate ssDNA C>T damage from dsDNA C>T mutations. dsDNA 'Blood, heat' samples were heat treated at 56 °C and 72 °C (both 3 hours and 6 hours for each). dsDNA 'Blood': $n = 4$ samples; dsDNA 'Kidney and liver': $n = 10$ samples. **a**,**c**,**d**, HiDEF-seq spectra were corrected for trinucleotide context opportunities. **b**, Bars and error bars show point estimates and their Poisson 95% confidence intervals, and statistical analysis was performed using two-sided Poisson rates ratio tests; from left to right, $^*P = 0.001$, 0.35 (not significant (NS)), $^*P < 10^{-15}$, $^*P < 10^{-15}$.

or 5-hydroxyhydantoin (uracil-species)[34,35]. When unrepaired, these lesions result in dsDNA C>T mutations[34]. We reasoned that HiDEF-seq may detect these ssDNA cytosine to uracil-species events with single-molecule fidelity despite their low levels (estimated by mass spectrometry at less than 1 per 1 million bases[36]), as damaged cytosines would be mis-sequenced as thymines due to nucleotide misincorporation by the sequencer polymerase.

We began by investigating the burden and pattern of ssDNA C>T calls in the blood DNA of individuals without cancer predisposition, as blood can be processed rapidly without potential post-mortem DNA damage. We also extracted the DNA with room temperature incubations to avoid heat-induced deamination[37]. Blood DNA had 2.0×10^{-8} ssDNA C>T calls per base (mean of $n = 9$ samples from $n = 5$ individuals; range 9.8×10^{-9}–3.1×10^{-8}), comprising on average 71% of these samples' ssDNA calls (Extended Data Fig. 9a and Supplementary Tables 2 and 3). This burden, which may have either been present in vivo or partly arisen during laboratory processing, suggests that there are less than 250 cytosine to uracil-species deaminated bases per cell in blood leukocytes. Our detection level of 1 event per 50 million bases is on par with the most sensitive mass spectrometry methods[36,38]—which cannot determine the sequence context of damaged bases—and provides a low background for studying cytosine deamination processes. Notably, the spectrum of the combined ssDNA calls of these blood samples, projected to central pyrimidine contexts, most closely resembled COSMIC[23] SBS30 (cosine similarity = 0.83) (Fig. 4a,c), a signature associated with cytosine oxidative deamination damage repaired by DNA glycosylases[18,39,40]. Surprisingly, G>T ssDNA calls, which would be expected due to the commonly oxidized base 8-oxoguanine, were very infrequent in these blood samples (average of 6% of ssDNA calls, 1.5×10^{-9} ssDNA calls per base; range 0–2.9×10^{-9}), possibly due to the sequencer polymerase correctly incorporating dC across from 8-oxoguanine.

Given the high sensitivity of HiDEF-seq's ssDNA C>T detection, we investigated the effect of heat, an important source of laboratory-based cytosine deamination artefacts (as most DNA extraction methods use heat)[37]. We profiled purified blood DNA after heat incubation at 56 °C

and 72 °C, each for 3 hours (h) and 6 h. While heat did not affect dsDNA mutation burdens, HiDEF-seq measured a significant increase in ssDNA calls (29-fold for 72 °C, 6 h treatment), specifically C>T calls (97% of calls), with increasing temperature and time (Fig. 4b and Supplementary Tables 2 and 3). This observation led us to profile all of the samples in this study except four (neurons of individual 5344 and 3 tumour samples) at least once with a room temperature DNA extraction (Methods and Supplementary Table 1). Notably, HiDEF-seq library preparation temperatures do not exceed 37 °C (Methods).

Across all of the healthy tissues and cell lines that we profiled, only sperm had a similarly high percentage of ssDNA calls that were C>T (average, 94%; Extended Data Fig. 9a). Sperm also had a higher ssDNA C>T burden than the other sample types (average, 1.4×10^{-7} C>T calls per base; Extended Data Fig. 9a). This suggests that these are also cytosine deamination events and that sperm DNA either undergoes more in vivo cytosine deamination than DNA of other tissues, or that it incurs this damage ex vivo before sperm purification from semen, during sperm purification or freezing, and/or during DNA extraction. To distinguish between these possibilities, we profiled non-sperm samples with the same processes used to freeze sperm and extract DNA from sperm, and we profiled additional sperm samples purified using filter chips that mimic physiological separation of motile sperm (Methods). The former did not produce an increase in ssDNA C>T burden, and the latter measured similar C>T burdens to the previous sperm samples that were purified by standard density gradient centrifugation (Supplementary Table 2 and Supplementary Note 5). These results suggest that sperm incur an elevated cytosine deamination burden either in vivo or ex vivo during the time (<1 h) that semen liquefies in the laboratory before sperm purification. In both cases, the elevated cytosine deamination burden would likely be present in sperm fertilizing the egg, and the egg's DNA repair machinery would then repair the damage[41]. Moreover, sperm ssDNA C>T calls did not exhibit transcription level or transcription strand biases (Supplementary Note 6).

Notably, all sperm and heat-treated blood DNA samples exhibited similar ssDNA C>T spectra, and the projection of these ssDNA spectra to dsDNA spectra again closely matched COSMIC dsDNA signature SBS30 (average cosine similarities of 0.92 and 0.95 for sperm and 72 °C heat samples, respectively) (Fig. 4c and Extended Data Fig. 9b). Using all of the above sperm and heat-damage samples, we next extracted this ssDNA signature, which we named SBS30ss* (cosine similarity = 0.94 to SBS30) (Fig. 4d). COSMIC signature SBS30 is associated with *NTHL1* and *UNG* biallelic loss-of-function mutations[18,39] and with formalin fixation[42]. *NTHL1* and *UNG* encode DNA glycosylases that initiate base excision repair of oxidized pyrimidines, including uracil-species resulting from cytosine oxidation[40]. Our finding that in vitro heating of purified DNA leads to a ssDNA damage signature, SBS30ss*, that matches the in vivo dsDNA SBS30 signature indicates that the SBS30ss* process is active in vivo, and that its pattern reflects the nucleotide context bias of the primary biochemical process of cytosine deamination, probably through an oxidized intermediate.

To further characterize the ssDNA C>T calls in heat-treated DNA and sperm, we took advantage of the single-molecule sequencer's polymerase kinetic data that record the duration of each nucleotide incorporation (pulse width (PW)) and the time between nucleotide incorporations (interpulse duration (IPD)) (Fig. 4e). PW and IPD encode unique kinetic signatures for different canonical and damaged bases[43]. ssDNA C>T calls in heat-treated DNA and sperm exhibited a distinct PW and IPD kinetic signature compared to dsDNA C>T mutations (for the mutation strand containing thymine) (Fig. 4f, Methods and Extended Data Fig. 9c,d,g). These results provide further evidence that the ssDNA C>T calls are uracil-species arising from cytosine deamination damage and exclude the possibility that they are cytosine to thymine changes. We further validated that nearly all ssDNA C>T calls in heat-treated DNA and sperm are uracil-species by incubating three of these HiDEF-seq libraries with uracil DNA glycosylase and endonuclease VIII. This eliminated the SBS30ss* pattern and nearly all ssDNA C>T calls (Supplementary Note 7 and Supplementary Tables 2 and 3).

We also evaluated heating of DNA in five different buffers and in water. Heating in water or Tris buffer without additional salt increased cytosine damage 66-fold relative to heating in higher-salt buffers, with slight differences in ssDNA C>T patterns (Extended Data Fig. 9e,f and Supplementary Table 2). As low salt decreases DNA duplex stability at elevated temperatures, these results suggest that the in vivo mechanism of SBS30ss*/SBS30 is cytosine deamination while DNA is transiently single-stranded.

Patterns of APOBEC3A-induced damage

HiDEF-seq's detection of cytosine deamination damage with single-molecule fidelity motivated us to define a ssDNA damage signature for APOBEC3A that was recently distinguished as the key contributor to cytosine deamination caused by APOBEC3 family proteins[44]. We expressed human APOBEC3A in primary human fibroblasts and extracted a ssDNA signature, which we named SBS2ss*, with strong similarity to APOBEC3A's associated COSMIC dsDNA signature SBS2 (cosine similarity = 0.92) (Extended Data Fig. 10a–f). Notably, SBS2ss* contained additional low-level peaks of ssDNA C>T calls outside the TCN contexts characteristic of SBS2 (Extended Data Fig. 10f and Supplementary Note 8). Moreover, the absence of any appreciable ssDNA C>A or C>G calls (Extended Data Fig. 10e,f) provides further strong evidence that the COSMIC SBS13 signature associated with APOBEC3A arises by base excision followed by error-prone translesion synthesis across the resulting abasic sites[44] (Supplementary Note 8).

Profiling the mitochondrial genome

Previous studies measured an approximately 20–40-fold higher dsDNA mutation rate with age in the mitochondrial genome than in the nuclear genome[15]. However, the mechanism by which the mitochondrial genome mutates remains unclear[45–48]. While it was long assumed to be primarily due to oxidative damage[47], recent studies instead support a mechanism linked to replication[45–49]. Specifically, A>G and C>T dsDNA mutations are highly enriched on the mitochondrial heavy (G+T-rich) strand, with a frequency that decreases with distance from the heavy strand origin of replication in the direction of heavy strand synthesis[45,46,48,49]. Several potentially overlapping hypotheses have been proposed for these findings: (1) strand-displacement replication leaves the heavy strand exposed longer as ssDNA, making it vulnerable to deamination of adenine and cytosine that are then mispaired during replication with cytosine and adenine, respectively[45,46,48]; (2) strand asymmetries in polymerase misincorporation of canonical nucleotides[46,47]; and (3) strand asymmetries in DNA repair[46]. Importantly, assuming that DNA repair is not substantially more efficient in mitochondria than in nuclei[50] and that most mutagenic mitochondrial ssDNA lesions can be detected by HiDEF-seq, then possibilities (2) and (3) should exhibit significantly higher HiDEF-seq ssDNA burdens in the mitochondrial genome than in the nuclear genome—since HiDEF-seq detects an increased ssDNA burden in CMMRD and *POLE* PPAP samples that have even lower dsDNA mutation rates than mitochondria (8.1-fold and 5.4-fold lower, respectively) (Fig. 5a and Extended Data Fig. 7d). However, possibility (1) would not yield a substantial difference in HiDEF-seq ssDNA burdens between the mitochondrial and nuclear genomes because HiDEF-seq would not capture denatured mitochondrial ssDNA in which the ssDNA damage events occur, and these ssDNA damage events would be rapidly transformed into dsDNA changes by replication. We investigated HiDEF-seq's mitochondrial dsDNA and ssDNA calls to assess these hypotheses.

We focused on liver and kidney samples, which yield more mitochondrial DNA (average 1% of sequenced molecules) than other tissues, and we also purified mitochondria from five liver samples to further

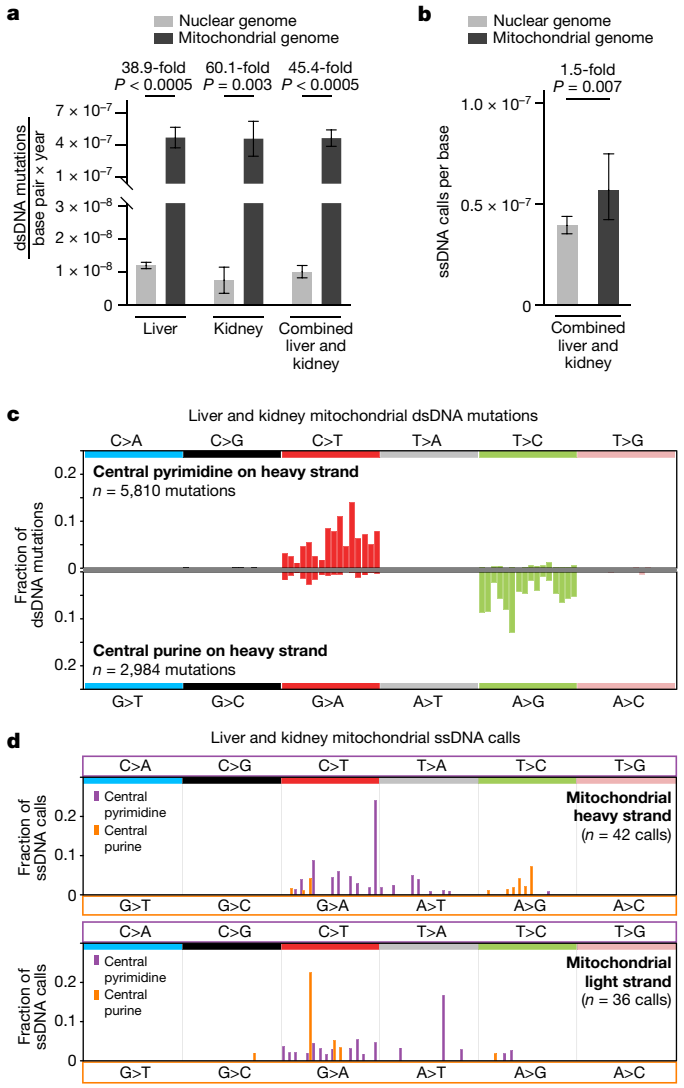


Fig. 5 | Mitochondrial genome dsDNA and ssDNA call burdens and patterns. **a**, Nuclear versus mitochondrial genome dsDNA mutation rates. Mitochondrial rates are from the regressions in Extended Data Fig. 11a, which were performed similarly for the nuclear genome and for liver and kidney samples combined. *P* values were calculated using analysis of variance (ANOVA) comparing two weighted least-squares linear regression models of mutation burden versus age and genome type covariates: one with and one without an 'age × genome type' interaction term (an estimate of the difference in dsDNA mutation rate depending on whether it is the nuclear or mitochondrial genome). **b**, ssDNA call burdens in the nuclear versus mitochondrial genomes after combining the calls of liver and kidney samples shown in Extended Data Fig. 11a, excluding from the nuclear genome burden the liver samples from which mitochondria were enriched as, due to low DNA inputs, these samples were profiled with HiDEF-seq with A-tailing, which induces ssDNA T>A artefacts in the nuclear genome of post-mortem liver. *P* value was calculated using a two-sided Poisson rates ratio test. **c**, dsDNA mutation spectrum, corrected for trinucleotide context opportunities, of the liver and kidney samples shown in Extended Data Fig. 11a for the mitochondrial genome heavy strand, separated by pyrimidine (top) and purine (bottom) contexts. **d**, Spectrum of mitochondrial ssDNA calls combined from the liver and kidney samples shown in Extended Data Fig. 11a plus all bulk (that is, non-mitochondria enriched) liver and kidney samples profiled by HiDEF-seq with A-tailing, as the ssDNA T>A artefact that A-tailing can incur in these post-mortem tissues (Supplementary Note 1) is orthogonal to the contexts of mitochondrial mutagenesis. Spectra are corrected for trinucleotide context opportunities, separately for each strand. Excluding bulk samples profiled by HiDEF-seq with A-tailing yields a similar spectrum (Extended Data Fig. 11c). **a**, Error bars show the 95% confidence intervals from regressions. **b**, Bars and error bars show point estimates and their Poisson 95% confidence intervals.

increase mitochondrial DNA yield (average of 13% of molecules; Supplementary Table 1). Mitochondrial dsDNA mutation rates measured by HiDEF-seq were 38.9- and 60.1-folder higher in liver and kidney, respectively, than the dsDNA mutation rates of the nuclear genomes of these tissues (Fig. 5a and Extended Data Fig. 11a). Combining liver and kidney samples, the difference was 45.4-fold (Fig. 5a). HiDEF-seq also detected the expected highly asymmetric pattern of A>G and C>T dsDNA mutations on the heavy strand, and the heavy strand's A>G mutation spectrum had strong similarity to SBS30ss* and SBS30 (both cosine similarities = 0.91) (Fig. 5c, Extended Data Fig. 11b and Supplementary Note 9).

Notably, despite the mitochondrial genome's significantly higher dsDNA mutation rate, its ssDNA call burden in liver and kidney was only 1.5-fold higher (95% confidence interval: 1.1–2.1) than the ssDNA call burden of the nuclear genome (Fig. 5b). While the number of mitochondrial ssDNA calls was low, these were concentrated in sequence contexts consistent with the dsDNA mutation spectrum (Fig. 5d, Extended Data Fig. 11c and Supplementary Note 9). Together, these data strengthen the evidence that the mitochondrial genome mutates primarily during replication, possibly through DNA damage on the heavy strand while it is single-stranded and, to a lesser extent, through cytosine deamination on the light strand (Supplementary Note 9).

Discussion

Profiling dsDNA mutations provides information on past mutational events, while profiling ssDNA mismatches and damage provides a real-time view of DNA lesions that reflects the current equilibrium between DNA damage, repair, and replication. Once ssDNA mismatches and damage transform into dsDNA mutations, information is lost about the originating lesions. This gap in studying mutagenesis motivated us to develop HiDEF-seq—a single-molecule sequencing approach that achieves single-molecule fidelity. Our approach opens new avenues for studying DNA damage and mutation processes.

Mutational signatures have transformed the study of cancer and mosaic mutations[4], but current signatures reflect only dsDNA mutations. Here we have begun to define ssDNA signatures, specifically: SBS10ss, SBS14ss, SBS30ss* and SBS2ss* (Supplementary Table 7). SBS10ss and SBS14ss arise from misincorporation of canonical (that is, non-damaged) nucleotides during replication. ssDNA mismatches of canonical nucleotides probably also occur outside the setting of replication. For example, signature SBS5 is ubiquitous in all cells, including post-mitotic neurons[8,51], and a recent study indicates that SBS5 may be caused by translesion polymerases[44]. This implies a mechanism of canonical nucleotide misincorporation that may become detectable by HiDEF-seq with higher-throughput instruments. We anticipate that HiDEF-seq will spur efforts to create a comprehensive catalogue of ssDNA signatures that complements the existing catalogue of dsDNA signatures. It will then be important to relate specific ssDNA and dsDNA signatures to each other, as these relationships will encode information about DNA damage, repair, and replication dynamics. Furthermore, as we have shown here, HiDEF-seq may be used to systematically assess potential damage caused by laboratory tissue and DNA processing.

The prevailing view that single-molecule sequencers have relatively high cost may have deterred their use in studying mosaic mutations and rare events, with the exception of in vitro polymerase and bacterial mutagenesis studies[52,53]. Since HiDEF-seq captures data from both DNA strands more efficiently than short-read duplex sequencing, it is only around 4.6-fold more expensive for dsDNA mosaic mutation detection than short-read duplex sequencing, and new sequencing instruments will reduce this to an approximately 2.8-fold difference (and about 1.6-fold for large-fragment HiDEF-seq) (Supplementary Note 10). One limitation of HiDEF-seq is that it does not achieve single-molecule

fidelity for insertions and deletions (indels) due to high sequencing error rates for these events in single-molecule sequencing[12]. This may become feasible with improved sequencing fidelity and indel-tuned consensus sequence calling[12]. Moreover, HiDEF-seq does not currently detect types of ssDNA damage that do not affect base pairing or that cannot be replicated by the sequencing polymerase. Since diverse types of ssDNA damage alter sequencing polymerase kinetics[43], other types of damage may be feasible to detect in the future with single-molecule fidelity.

The high mutation rates of CMMRD and PPAP syndromes put their abnormal ssDNA call burdens and patterns within range of currently feasible single-molecule sequencing depth. However, we did not detect altered ssDNA burdens or patterns in cancer predisposition syndromes involving nucleotide excision repair or base repair, probably due to current limitations of sequencing depth and/or their mutational mechanisms involving types of ssDNA damage that we do not currently detect. We anticipate that future higher-throughput single-molecule sequencing combined with kinetics analyses will reveal additional ssDNA signatures in other cancer predisposition syndromes and in individuals with normal mutation rates.

Diverse methods profile DNA damage by enzymatic alteration at damage sites or by affinity enrichment, but their lack of single-molecule fidelity yields low-resolution damage patterns[10]. HiDEF-seq's single-molecule fidelity for cytosine deamination damage revealed SBS30ss*. In healthy tissues, we detect SBS30ss* but not an SBS1ss* signature corresponding to SBS1, suggesting that SBS30ss* in healthy tissues reflects primarily ex vivo cytosine deamination that obscures in vivo SBS1ss* (Supplementary Note 11). However, in sperm, the higher burden of SBS30ss* may reflect in vivo cytosine deamination that accumulates in the absence of effective DNA repair and is later repaired after fertilization[41]. Nevertheless, when SBS1 is elevated, HiDEF-seq can detect its ssDNA precursors (Supplementary Note 4).

HiDEF-seq may also find utility in experimental systems to dissect the kinetics of the DNA damage, repair, and replication equilibrium—for example, combined with in vitro genetic and other manipulations, with synchronization of the cell cycle, and in reconstituted enzyme systems. Sequencing single-strand changes in DNA with single-molecule fidelity will greatly advance our understanding of the origins of mutations.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07532-8.

1. Mustjoki, S. & Young, N. S. Somatic mutations in "benign" disease. *N. Engl. J. Med.* **384**, 2039–2052 (2021).
2. Vijg, J. & Dong, X. Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* **182**, 12–23 (2020).
3. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat. Rev. Genet.* **22**, 672–686 (2021).
4. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* **21**, 619–637 (2021).
5. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
6. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
7. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
8. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
9. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508 (2012).
10. Sloan, D. B., Broz, A. K., Sharbrough, J. & Wu, Z. Detecting rare mutations and DNA damage with sequencing-based methods. *Trends Biotechnol.* **36**, 729–740 (2018).
11. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
12. Baid, G. et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2022).
13. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
14. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
15. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
16. 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. USA* **118**, e2013106118 (2021).
17. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**, 1434–1442 (2021).
18. Zou, X. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* **2**, 643–657 (2021).
19. Sanders, M. A. et al. Life without mismatch repair. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.14.437578 (2021).
20. Yurchenko, A. A. et al. XPC deficiency increases risk of hematologic malignancies through mutator phenotype and characteristic mutational signature. *Nat. Commun.* **11**, 5834 (2020).
21. Robinson, P. S. et al. Inherited MUTYH mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *Nat. Commun.* **13**, 3949 (2022).
22. Lujan, S. A., Williams, J. S. & Kunkel, T. A. DNA polymerases divide the labor of genome replication. *Trends Cell Biol.* **26**, 640–654 (2016).
23. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
24. Lujan, S. A. et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.* **24**, 1751–1764 (2014).
25. Shinbrot, E. et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).
26. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
27. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
28. Bulock, C. R., Xing, X. & Shcherbakova, P. V. Mismatch repair and DNA polymerase δ proofreading prevent catastrophic accumulation of leading strand errors in cells expressing a cancer-associated DNA polymerase ϵ variant. *Nucleic Acids Res.* **48**, 9124–9134 (2020).
29. Shlien, A. et al. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat. Genet.* **47**, 257–262 (2015).
30. Hodel, K. P. et al. Explosive mutation accumulation triggered by heterozygous human Pol ε proofreading-deficiency is driven by suppression of mismatch repair. *eLife* **7**, e32692 (2018).
31. Haradhvala, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).
32. Hodel, K. P. et al. POLE mutation spectra are shaped by the mutant allele identity, its abundance, and mismatch repair status. *Mol. Cell* **78**, 1166–1177 (2020).
33. Kunkel, T. A. & Erie, D. A. Eukaryotic mismatch repair in relation to DNA replication. *Ann. Rev. Genet.* **49**, 291–313 (2015).
34. Shinmura, K. et al. Defective repair capacity of variant proteins of the DNA glycosylase NTHL1 for 5-hydroxyuracil, an oxidation product of cytosine. *Free Radic. Biol. Med.* **131**, 264–273 (2019).
35. Dizdaroglu, M. Oxidatively induced DNA damage and its repair in cancer. *Mutat. Res. Rev. Mutat. Res.* **763**, 212–245 (2015).
36. Madugundu, G. S., Cadet, J. & Wagner, J. R. Hydroxyl-radical-induced oxidation of 5-methylcytosine in isolated and cellular DNA. *Nucleic Acids Res.* **42**, 7450–7460 (2014).
37. Chen, G., Mosier, S., Gocke, C. D., Lin, M.-T. & Eshleman, J. R. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol. Diagn. Ther.* **18**, 587–593 (2014).
38. Tretyakova, N., Villalta, P. W. & Kotapati, S. Mass spectrometry of structurally modified DNA. *Chem. Rev.* **113**, 2395–2436 (2013).
39. Grolleman, J. E. et al. Mutational signature analysis reveals NTHL1 deficiency to cause a multi-tumor phenotype. *Cancer Cell* **35**, 256–266 (2019).
40. Krokan, H. E. & Bjørås, M. Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5**, a012583 (2013).
41. Stringer, J. M., Winship, A., Liew, S. H. & Hutt, K. The capacity of oocytes for DNA repair. *Cell. Mol. Life Sci.* **75**, 2777–2792 (2018).
42. Guo, Q. et al. The mutational signatures of formalin fixation on the human genome. *Nat. Commun.* **13**, 4487 (2022).
43. Clark, T. A., Spittle, K. E., Turner, S. W. & Korlach, J. Direct detection and sequencing of damaged DNA bases. *Genome Integr.* **2**, 10 (2011).
44. Petljak, M. et al. Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
45. Sanchez-Contreras, M. et al. A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. *Nucleic Acids Res.* **49**, 11103–11118 (2021).
46. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, e02935 (2014).

47. Kauppila, J. H. K. & Stewart, J. B. Mitochondrial DNA: radically free of free-radical driven mutations. *Biochim. Biophys. Acta* **1847**, 1354–1361 (2015).
48. Kennedy, S. R., Salk, J. J., Schmitt, M. W. & Loeb, L. A. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794 (2013).
49. Yuan, Y. et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
50. Fontana, G. A. & Gahlon, H. L. Mechanisms of replication and repair in mitochondrial DNA deletion formation. *Nucleic Acids Res.* **48**, 11244–11258 (2020).
51. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
52. Matsuda, T., Matsuda, S. & Yamada, M. Mutation assay using single-molecule real-time (SMRTTM) sequencing technology. *Genes Environ.* **37**, 15 (2015).
53. Hestand, M. S., Houdt, J. V., Cristofoli, F. & Vermeesch, J. R. Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutat. Res.* **784–785**, 39–45 (2016).

Methods

Sample sources

Post-mortem tissues obtained by the NIH NeuroBioBank (University of Maryland site) were frozen in isopentane-liquid nitrogen baths and stored at −80 °C until use. Post-mortem tissues obtained by the International Replication Repair Deficiency Consortium (IRRDC) biobank were frozen and stored at −80 °C until use. Blood was obtained from individuals enrolled in human subjects research of the New York University Grossman School of Medicine, the IRRDC, the University of Pittsburgh and the Cryos International Sperm Bank. All blood samples were collected in EDTA tubes and frozen immediately after collection until use. Tumour samples were obtained from the IRRDC and were frozen and stored at −80 °C until use. Semen samples (processing details described in the 'Sperm purification' section) were obtained at Cryos International Sperm Bank from individuals enrolled in human subjects research approved by the New York University Grossman School of Medicine Institutional Review Board, except for participants D1 and D2, who were enrolled in human subjects research conducted by Cryos International Sperm Bank. Lymphoblastoid cell lines were obtained from Coriell Institute and the IRRDC. Primary fibroblasts were obtained from Coriell Institute and the IRRDC. All of the samples were collected under human subjects research protocols approved by either the New York University Grossman School of Medicine Institutional Review Board, the Hospital for Sick Children (SickKids) Research Ethics Board as part of the IRRDC, the Cryos International Sperm Bank scientific advisory committee or the University of Pittsburgh Institutional Review Board.

The source, sex, age at collection, and post-mortem interval of each sample are provided in Supplementary Table 1.

Sperm purification

After collection at the Cryos International Sperm Bank, semen underwent liquefaction at room temperature for 30 to 60 min. Semen then immediately underwent initial purification for sperm using density gradient centrifugation followed by a wash with HEPES-buffered medium[54]. For semen from individuals D1 and D2, sperm were purified from half of each semen sample using this method, and sperm were purified from the other half using the ZyMot Multi (850 µl) Sperm Separation Device (ZyMot) according to the manufacturer's instructions. After addition of cryopreservation media, sperm were stored in liquid nitrogen until further use.

Cryopreserved sperm that previously underwent initial purification by density gradient centrifugation were further purified in the laboratory with a second density gradient centrifugation and two additional washes, as follows. First, the following reagents were equilibrated to room temperature: ORIGIO gradient 40/80 buffer (Cooper Surgical, 84022010), Origio sperm wash buffer (Cooper Surgical, 84050060) and Quinn's Advantage sperm freezing medium (Cooper Surgical, ART-8022). In a 15 ml tube, 1 ml of Origio 80 buffer was placed at the bottom, and 1 ml of Origio 40 buffer was gently layered on top. Sperm were thawed at room temperature for 15 min, gently mixed with a pipette, and carefully layered on top of the Origio 40 buffer. The tube was then centrifuged in a swinging-bucket centrifuge at 400*g* for 20 min at room temperature with low acceleration and deceleration speeds. The supernatant was aspirated, leaving 500 µl of sperm/buffer at the bottom. The sperm was transferred to a new 15 ml tube and diluted with 5 ml sperm wash buffer. The tube was mixed by inverting ten times and centrifuged in a swinging-bucket centrifuge at 300*g* for 10 min at room temperature with maximum acceleration and deceleration. The supernatant was removed, leaving about 350 µl of sperm/buffer at the bottom. The sperm was then washed again in the same way with 5 ml of sperm wash buffer, and the supernatant was removed, leaving about 250 µl of sperm/buffer at the bottom of the tube. After pipette mixing, an aliquot of this sperm was transferred to a 2 ml DNA LoBind microtube (Eppendorf) for immediate DNA extraction and general evaluation using a haemocytometer. The remaining sperm was diluted dropwise with a 1:1 volumetric ratio of sperm freezing medium, incubated at room temperature for 3 min, frozen in a Mr. Frosty freezing container (Thermo Fisher Scientific) in a −80 °C freezer for 24 h and then transferred to a liquid nitrogen freezer.

Cerebral cortex neuronal nuclei purification

Cerebral cortex neuronal nuclei were isolated as previously described[5] from post-mortem frontal cortex (Brodmann area 9, left hemisphere) of individuals who did not have any known neurological or psychiatric disease. Specifically, approximately 1 g of frozen tissue from each individual was cut into 5 mm^3 pieces and added to 9 ml of chilled lysis buffer (0.32 M sucrose, 10 mM Tris HCl pH 8, 5 mM $CaCl_2$, 3 mM magnesium acetate, 0.1 mM EDTA, 1 mM DTT, 0.1% Triton X-100) in a large dounce homogenizer (Sigma-Aldrich, D9938). While on ice, the tissue was dounced 20 times each with pestle size A and then B. The homogenate was layered on a 7.4 ml sucrose cushion (1.8 M sucrose, 10 mM Tris HCl pH 8, 3 mM magnesium acetate, 1 mM DTT) in an ultracentrifuge tube on ice. The tubes were centrifuged (Thermo Fisher Scientific, Sorvall LYNX 6000) at 10,000 rpm for 1 h at 4 °C. The resulting supernatant was removed, and 500 µl of nuclei resuspension buffer (3 mM $MgCl_2$ in 1× phosphate-buffered saline) was added on top of the pellet and then incubated on ice for 10 min. The pellet was then gently resuspended. Antibody staining buffer was prepared by adding 1.2 µg of NeuN-Alexa-647 (Abcam, ab190565) to 400 µl of antibody staining buffer (3% BSA in nuclei resuspension buffer) and inverted gently to mix. Then, 400 µl of antibody staining buffer was added to 1 ml of nuclei and the sample was rotated at 4 °C for 30 min. NeuN-positive nuclei were gated as shown in Supplementary Note 12. NeuN-positive nuclei were collected in 30 µl of nuclei buffer in 1.5 ml LoBind tubes (Eppendorf) by fluorescence-activated nuclei sorting on a SONY LE-SH800 sorter. After sorting, a 1:1 volumetric ratio of 80% glycerol was added to sorted nuclei for a final concentration of 40% glycerol to stabilize nuclei during centrifugation. Nuclei were centrifuged at 4 °C, 500*g* for 10 min. The supernatant was removed and nuclei pellets were immediately frozen at −80 °C.

Extraction and isolation of mitochondria

Mitochondria were extracted and isolated from 300–500 mg of tissue using the Mitochondria Extraction Kit (Miltenyi Biotec) and Mitochondria Isolation Kit (Miltenyi Biotec), according to the manufacturer's Extraction Kit protocol, with the following modifications: (1) protease inhibition buffer was prepared with 100× HALT protease inhibitor cocktail (Thermo Fisher Scientific); (2) minced tissue was resuspended with a larger 2 × 2.5 ml volume of protease inhibitor buffer instead of 2 × 1 ml; (3) after homogenization, the homogenate was passed through a 30 µm SmartStrainer (Miltenyi Biotec); (4) the SmartStrainer was washed with 2 × 2.5 ml of solution 3 instead of 2 × 1 ml; (5) before adding TOM22 antibody, the homogenate was diluted with Separation Buffer to a volume of 25 ml instead of 10 ml; and (6) 125 µl of TOM22 antibody was used per sample instead of 50 µl. Final mitochondria pellets were frozen at −20 °C for subsequent DNA extraction.

Cell culture for direct profiling

Lymphoblastoid cell lines were cultured at 37 °C, 5% CO_2, and ambient oxygen in T25 flasks with RPMI 1640 medium (Thermo Fisher Scientific, 61870036) supplemented with 15% fetal bovine serum and penicillin–streptomycin. Cells were passaged to new medium approximately every 2–3 days.

Fibroblasts were cultured at 37 °C, 5% CO_2 and ambient oxygen in T25 flasks with DMEM medium (Thermo Fisher Scientific, 10569010) supplemented with 10% fetal bovine serum and penicillin–streptomycin. Cells were passaged to new medium every 3–5 days before reaching full confluency. Cells were collected for DNA extraction at 80–90% confluency using trypsin-EDTA.

For the experiment testing the potential effect of sperm freezing medium on cytosine deamination, we resuspended the collected pellet of fibroblasts in Origio sperm wash buffer, mixed with a 1:1 volume ratio of Freezing Medium TYB with Glycerol & Gentamicin (Irvine Scientific), and froze the cells in a Mr. Frosty container (Thermo Fisher Scientific) at −80 °C followed by transfer to a liquid nitrogen freezer. After thawing, cells were either washed once with PBS followed by Puregene DNA extraction or they were processed using the same method of DNA extraction that was used for sperm (the details of each method are described in the 'DNA extraction' section).

Lentivirus experiments

Lentivirus plasmid design and synthesis. The lentivirus transfer plasmid design and sequences are listed in Supplementary Table 8. APOBEC3A constructs included a human gamma globin intron 2 sequence to prevent expression of the mutagenic protein during bacterial cloning[55]. Gene inserts were synthesized and cloned by GenScript into a pLVX-TetOne lentiviral vector (Takara). The pLVX-TetOne vector was used to enable temporal control of gene expression using doxycycline. This prevents expression of encoded mutagenic proteins during lentiviral packaging, which could mutate the lentiviral transfer plasmid and lentiviral RNA to create non-functional lentiviruses. GenScript verified gene inserts by sequencing and prepared quality-controlled quantities of transfer plasmid sufficient for lentiviral packaging.

Lentivirus packaging. Lenti-X 293T cells (Takara) were cultured at 37 °C, 5% CO_2 and ambient oxygen in T75 collagen-coated flasks (ZenBio) with DMEM medium (Thermo Fisher Scientific, 11995065) supplemented with 10% tetracycline-free fetal bovine serum (Takara). Cells were transfected at about 80% confluency. The lentiviral packaging transfection mix was prepared by combining 0.8 ml DMEM (Thermo Fisher Scientific; 11995065), 20 µl pC-Pack2 second-generation lentiviral packaging plasmid mix (Cellecta, CPCP-K2A), lentiviral transfer plasmid (10 µg for eGFP plasmid; 12.5 µg for APOBEC3A plasmids), and 36 µl PureFection transfection reagent (System Biosciences). Note that a second-generation packaging system was necessary because fourth-generation packaging systems contain a Tet-Off gene that would cause the pLVX-TetOne gene insert to be expressed during packaging, and third-generation packaging systems do not contain the *tat* gene required for efficient packaging of the fourth-generation pLVX-TetOne transfer plasmid. Cells were transfected by adding this transfection mix to cells in fresh 10 ml of the above medium. The next day, an additional 8 ml of the above medium was added to the cells. Then, 72 h after transfection, the cell medium was collected and centrifuged at 500*g* for 10 min to pellet the cell debris. The ~18 ml supernatant was mixed with 6 ml of Lenti-X Concentrator (Takara), incubated for at least 3 h at 4 °C and centrifuged at 1,500*g* for 45 min at 4 °C. The lentivirus pellet was resuspended in DMEM medium (Thermo Fisher Scientific, 10569010) supplemented with 10% standard fetal bovine serum and penicillin–streptomycin. Aliquots of lentivirus were flash-frozen in liquid nitrogen and stored at −80 °C.

Lentiviral particles were quantified after thawing using Lenti-X GoStix Plus (Takara). The resulting GoStix values were multiplied by 1.25×10^7 to obtain the lentiviral particle per ml concentration.

Lentivirus transduction. Fibroblasts were cultured at 37 °C, 5% CO_2 and ambient oxygen in T75 flasks with DMEM medium (Thermo Fisher Scientific, 10569010) supplemented with 10% fetal bovine serum and penicillin–streptomycin. Cells were transduced with lentivirus at about 60% confluency in 15 ml of the above medium supplemented with 8 µg ml^{-1} polybrene (Sigma-Aldrich, H9268). The amount of lentivirus added was calculated as follows: ([estimated 900,000 cells in a 60% confluent T75 flask] × [250 infectious units per cell])/([previously measured concentration of lentiviral particles per ml]/[estimated 100 viral particles per infectious unit]). The factor of 250 infectious units per cell was optimized to obtain > 80% GFP-positive cells using the eGFP lentivirus. Then, 16 h after transduction, the medium was replaced with a new 15 ml of the above medium (without polybrene) supplemented with 250 ng ml^{-1} doxycycline (Takara, 631311). After an additional 48 h, the medium was replaced with a new 15 ml of the above doxycycline medium. After an additional 24 h, cells were collected for DNA extraction using trypsin-EDTA.

DNA extraction

The DNA-extraction method used for each sample is listed in Supplementary Table 1. Details of each DNA extraction method are provided below.

DNA extraction from sperm for HiDEF-seq. An aliquot of washed sperm (that is, after the washes that are performed after density gradient centrifugation) was centrifuged at 300*g* for 5 min at room temperature. The supernatant was removed, leaving approximately 50 µl of sperm/buffer at the bottom of the microtube. The tube was tapped gently five times to break up the sperm pellet before adding lysis buffer.

If starting with frozen sperm instead of an aliquot of washed sperm, the frozen sperm vial was rapidly thawed in a 37 °C water bath, gently mixed with a pipette, and an aliquot was transferred to a 2 ml DNA LoBind microtube for DNA extraction. The remaining sperm was frozen again. The DNA extraction aliquot was diluted with 600 µl of Origio sperm wash buffer, centrifuged at 300*g* for 5 min at room temperature, and the supernatant was removed to leave approximately 100 µl of sperm/buffer at the bottom. The sperm was diluted again with 600 µl of Origio sperm wash buffer, centrifuged at 300*g* for 5 min at room temperature, and the supernatant was removed to leave approximately 50 µl of sperm/buffer at the bottom. The tube was tapped gently five times to break up the sperm pellet before adding lysis buffer.

Sperm DNA extraction was based on a previous study[56], with some modifications, including optimizations we performed that showed that tris(2-carboxyethyl)phosphine (TCEP) can be reduced from 50 mM to 2.5 mM in the lysis buffer. Specifically, sperm lysis buffer was prepared by combining (for each sample) 497.5 µl of Qiagen Buffer RLT (Qiagen) without β-mercaptoethanol and 2.5 µl of 0.5 M Bond-Breaker TCEP Solution (Thermo Fisher Scientific) for a lysis buffer with 2.5 mM TCEP final concentration. Then, 500 µl of sperm lysis buffer and 100 mg of 0.2 mm stainless-steel beads (Next Advance, SSB02-RNA) were added without mixing to each sample. Homogenization was then performed using a TissueLyser II instrument (Qiagen) at 20 Hz for 4 min (samples profiled by HiDEF-seq without nick ligation: SPM-1004, SPM-1020; samples profiled by HiDEF-seq with nick ligation and A-tailing, and samples profiled by NanoSeq: SPM-1002, SPM-1004, SPM-1013, SPM-1020; samples profiled by HiDEF-seq with nick ligation in large fragments: SPM-1002, SPM-1020) or 30 s (samples profiled by HiDEF-seq with nick ligation and A-tailing: SPM-1060, D1, D2; sample profiled by HiDEF-seq with nick ligation without A-tailing: SPM-1013; sample profiled by NanoSeq: SPM-1060; and samples profiled by HiDEF-seq with nick ligation and with uracil DNA glycosylase/endonuclease VIII treatment: SPM-1002 and SPM-1004). DNA was then extracted from the lysate using the QIAamp DNA Mini Kit (Qiagen) with a modified protocol as follows. A 500 µl volume of buffer AL was added to each lysate and vortexed well. Then, 500 µl of 100% ethanol was added and vortexed well. The mixture was then applied to a QIAamp DNA Mini spin column and the remaining standard QIAamp protocol was followed. DNA was eluted with 100 µl of 10 mM Tris pH 8. RNase treatment was then performed by adding 12 µl of 10× PBS pH 7.4 (Gibco), 2 µl of Monarch RNase A (New England Biolabs (NEB)) and 6 µl nuclease-free water (NFW). The reaction was incubated at room temperature for 5 min and immediately purified using a 0.8× beads to sample volume ratio of SPRI beads (solid-phase reversible immobilization; made by

washing 1 ml Sera-Mag carboxylate-modified SpeedBead (Cytiva, 65152105050250) and resuspending the beads in 50 ml of 18% PEG-8000, 1.75 M NaCl, 10 mM Tris pH 8, 1 mM EDTA, 0.044% Tween-20). DNA was eluted from beads with 35 µl of 10 mM Tris/0.1 mM EDTA pH 8. For the experiments in which we processed previously extracted blood DNA and primary fibroblast DNA with the same process used for sperm DNA extraction, we inputted previously extracted DNA and followed the same process above beginning with addition of lysis buffer, with a homogenization time of either 30 s or 4 min with concordant results (Supplementary Table 2).

A somatic cell contamination assay was adapted from a previous study[57] and performed on all extracted sperm DNA samples to further confirm sperm purity. This assay amplifies four loci from bisulfite-treated DNA: three loci that are methylated in sperm but not in somatic cells (PCR7, PCR11, PCR31) and 1 locus that is methylated in somatic cells but not in sperm (PCR12). After bisulfite treatment and PCR amplification of each locus, the PCR amplicon is cut by a restriction enzyme only if the original DNA was methylated. Thus, this assay can detect somatic cell contamination. In total, 350 ng of each extracted sperm DNA and 350 ng of control human NA12878 lymphoblastoid cell line genomic DNA (Coriell Institute) were bisulfite-converted using the Zymo EZ DNA Methylation Kit (Zymo Research). The loci were amplified by PCR using the following primer sets: PCR7 (GGGTTATATGA TAGTTTATAGGGTTATT and TCTATTACTACCACTTCCTAAATCAA), PCR11 (TGAGATGTTTGTTAGTTTATTATTTTGG and TCATCTTCTC CCACCAAATTTC), PCR12 (TAGAGGGTAGTTTTTAAGAGGG and ATTAACCAACCTCTTCCATATTCTT) and PCR31 (TTTTAGTTTTGG GAGGGGTTGTTT and CTACCAAAATTAAAAACCAACCCAC). The PCR reaction contained 1.5 µl of bisulfite-converted DNA, 10 µl of 2× ZymoTaq PCR Mix (Zymo Research), PCR primers, and NFW to a final volume of 20 µl. The PCR reactions were optimized to contain the following final concentrations of each forward and reverse primer: 0.6 µM for PCR7 primers, 0.6 µM for PCR11 primers, 0.3 µM for PCR12 primers, and 0.45 µM for PCR31 primers. The PCR reactions were cycled as follows: 95 °C for 10 min; 40 cycles of 94 °C for 30 s, *X* °C for 30 s and 72 °C for 30 s; 72 °C for 7 min; and hold at 4 °C, where *X* (annealing temperature) was 49 °C for PCR7 and PCR11, 51 °C for PCR12 and 55 °C for PCR31. PCR reactions were purified by 2× volumetric ratio SPRI beads cleanup and eluted in 22 µl of 10 mM Tris pH 8. Restriction digests were performed by combining 5 µl of purified PCR product, restriction enzyme (10 units of HpyCH4IV (NEB) for PCR7 and PCR31, and 20 units of TaqI-v2 (NEB) for PCR11 and PCR12), 1 µl of 10× CutSmart buffer (NEB), and NFW for a total reaction volume of 10 µl. Restriction digestions were performed at 37 °C (HpyCH4IV) or 65 °C (TaqI-v2) for 60 min. Control reactions without restriction enzyme were performed for each sample/locus combination. A total of 5 µl of each restriction digest reaction was combined with 1 µl 6× TriTrack DNA loading dye (Thermo Fisher Scientific) and run on a 2% agarose gel prestained with ethidium bromide, followed by imaging of the gel.

DNA extraction from solid tissues for HiDEF-seq. Approximately 50–300 mg of tissue was cut in a Petri dish on dry ice and minced with a scalpel, followed by one of the following DNA-extraction methods, as specified for each sample in Supplementary Table 1.

Nucleobond HMW, MagAttract HMW, QIAamp. In this method, DNA was extracted and purified with three serial kits to maximize DNA purity. DNA was extracted using the NucleoBond HMW DNA Kit (Takara) according to the manufacturer's instructions with a 50 °C proteinase K incubation for 4.5 h. The eluted DNA was then further purified with the MagAttract HMW DNA Kit (Qiagen) according to the manufacturer's whole-blood purification protocol, except with proteinase K/RNase A incubation occurring at 56 °C for 20 min. The eluted DNA was then further purified using the QIAamp DNA Mini Kit (Qiagen) by diluting the DNA to a final volume of 200 µl and final 1× PBS concentration, adding 20 µl of proteinase K (Qiagen) and continuing according to the manufacturer's body fluids DNA purification protocol with a 56 °C proteinase K incubation for 10 min without RNase A treatment.

MagAttract HMW. We used the MagAttract HMW DNA Kit (Qiagen) according to the manufacturer's protocol for tissue, with a 2 h proteinase K digestion at 56 °C. DNA was eluted with 10 mM Tris pH 8.

Puregene. Tissue was pulverized inside a microtube while in a liquid-nitrogen cooled mini mortar and pestle (Bel-Art). DNA was then extracted using the Puregene DNA Kit (Qiagen) according to the manufacturer's protocol for tissues, except (1) the lysis step with proteinase K was performed at room temperature on a ThermoMixer C instrument (Eppendorf) at 1,400 rpm for 1 h; (2) the RNase A treatment was performed at room temperature for 20 min; and (3) the final DNA pellet was resuspended in 10 mM Tris pH 8 at room temperature for 1 h.

DNA extraction from cerebral cortex neuronal nuclei for HiDEF-seq. DNA was extracted from nuclei pellets using two methods, as specified for each sample in Supplementary Table 1.

QIAamp: we used the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's protocol, with lysis performed by adding 180 µl of buffer ATL and 20 µl of proteinase K to the nuclei pellet, followed by a 56 °C incubation for 1 h, and including RNase A treatment.

MagAttract: we used the MagAttract HMW DNA Kit according to the manufacturer's protocol for blood, after resuspending nuclei with 200 µl of 1× PBS, with a 30 min proteinase K digestion at room temperature.

DNA extraction from mitochondria for HiDEF-seq. DNA was extracted from mitochondria pellets using the Puregene DNA Kit (Qiagen) according to the manufacturer's protocol for tissues, except (1) the lysis step used 200 µl Cell Lysis Solution and 1.5 µl proteinase K and was performed at room temperature for 30 min; (2) the RNase A treatment was performed at room temperature for 20 min; and (3) the final DNA pellet was resuspended in 10 mM Tris pH 8 at room temperature without an extended incubation.

Note that, due to the relatively low yields of mitochondria DNA preparations, these samples were profiled with HiDEF-seq with A-tailing (see the 'HiDEF-seq library preparation' section).

DNA extraction from blood, lymphoblastoid cells, and fibroblasts for HiDEF-seq and germline sequencing. DNA from blood, lymphoblastoid cells, and fibroblasts (the latter two after resuspending cell pellets in 1× PBS)—except for blood from individuals whose tumours were profiled, fibroblasts testing the effect of sperm freezing medium, and fibroblasts from lentivirus experiments—was extracted using the MagAttract HMW DNA Kit according to the manufacturer's whole-blood purification protocol, with proteinase K incubation at room temperature for 30 min.

DNA from fibroblasts frozen in sperm-freezing medium and fibroblasts in lentivirus experiments was extracted using the Puregene DNA Kit according to the manufacturer's protocol for cultured cells, except (1) the protocol volumes were scaled 2.8-fold; (2) the lysis step used 840 µl cell lysis solution and 4.2 µl proteinase K and was performed at room temperature for 30 min; (3) the RNase A treatment was performed at room temperature for 20 min; and (4) the final DNA pellet was resuspended in 10 mM Tris pH 8 at 4 °C for 1 h.

We also performed an experiment that excluded a measurable cytosine deamination effect by possible leached iron from MagAttract magnetic beads (Extended Data Fig. 9e) by extracting an additional aliquot of DNA from the blood of individual 1901 using the Puregene DNA Kit according to the manufacturer's protocol for 'whole blood or bone marrow', except (1) 200 µl blood was first diluted with 100 µl of 1× PBS; (2) the cell lysis step was performed at room temperature; (3) the RNase A treatment was performed at room temperature for 20 min; and (4) the final DNA pellet was resuspended in 10 mM Tris pH 8 at 4 °C for 1 h.

DNA extraction from tumours and those individuals' corresponding blood for Illumina tumour and germline sequencing. DNA was extracted from tumours by first homogenizing the tumour using the Precellys 24 Tissue Homogenizer followed by the DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's protocol for animal tissues with a 56 °C incubation for 10 min. For individuals whose tumours were profiled, DNA was extracted from blood of those individuals using the PAXgene Blood DNA Kit (Qiagen) according to the manufacturer's protocol.

DNA extraction from saliva for Illumina germline sequencing. DNA was extracted using the QIAamp DNA Mini Kit according to the manufacturer's 'DNA purification from blood or body fluids' protocol and including RNase A treatment.

DNA extraction from the liver and spleen for Illumina germline sequencing. DNA of all of the samples was extracted using the QIAamp DNA Mini Kit according to the manufacturer's 'DNA purification from tissues' protocol with a 2 h proteinase K digestion at 56 °C and including RNase A treatment, except for liver of individual 5309, from which DNA was extracted using the MagAttract HMW DNA Kit according to the manufacturer's 'Fresh or Frozen Tissue' protocol with a 2 h proteinase K digestion at 56 °C.

DNA extraction from blood for Pacific Biosciences germline sequencing. DNA was extracted using the Chemagic DNA Blood 2k Kit (Perkin Elmer, CMG-1097) on the Chemagic 360 automated nucleic extraction instrument (Perkin Elmer) according to the manufacturer's protocols for DNA isolation from whole blood.

DNA quantity and quality measurements and storage. The concentration and quality of all DNA samples were measured using a NanoDrop instrument (Thermo Fisher Scientific), a Qubit 1× dsDNA HS Assay Kit (Thermo Fisher Scientific) and a Genomic DNA ScreenTape TapeStation Assay (Agilent). DNA was stored at −20 °C.

Illumina germline and tumour library preparation and sequencing

Illumina germline and tumour sequencing libraries were prepared using the TruSeq DNA PCR-Free Kit (Illumina) for all samples. At least 110 Gb (~36× genome coverage) of 150 bp paired-end sequencing per sample was obtained using a NovaSeq 6000 instrument (Illumina) by Psomagen, except for tumour sequencing and those individuals' corresponding germline sequencing, for which HiSeqX and NovaSeq 6000 instruments were used at the Centre for Applied Genomics at the Hospital for Sick Children.

Pacific Biosciences germline library preparation and sequencing

A total of 15 µg of DNA was cleaned up with 1× AMPure PB beads (Pacific Biosciences) and sheared to a target size of 14 kb using the Megaruptor 3 instrument (Diagenode) using the following settings: speed, 36; volume, 300 µl; concentration, 33 ng µl^{-1}. Library preparation was performed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) according to the manufacturer's instructions. Library fragments longer than 10 kb were selected using a PippinHT instrument (Sage Science). Size-selected libraries were sequenced on the Pacific Biosciences Sequel IIe system using the Sequel II Binding Kit 2.0 and Sequel II Sequencing Kit 2.0 (Pacific Biosciences), Sequencing primer v4 (Pacific Biosciences), 1 h binding time, 2 h pre-extension, adaptive loading, 2 h immobilization time, and 30 h movies.

Heat damage of DNA

DNA was heated in a volume of 62 µl at the temperature, for the time, and in the buffer listed for each sample in Supplementary Table 1, followed by incubation on ice up to a total of 6 h if the heating time was less than 6 h. Untreated samples in these experiments were incubated on ice for 6 h. The DNA was then input into HiDEF-seq library preparation.

NanoSeq library preparation and sequencing

NanoSeq libraries were prepared as previously described[8] with 50 ng DNA input from the same DNA aliquots used for HiDEF-seq.

HiDEF-seq library preparation and sequencing

Choice of restriction enzymes for DNA fragmentation. We performed in silico digests of the CHM13 v.1.0 human reference genome sequence[58] to identify restriction enzymes that (1) maximize the percentage of the genome between 1 and 4 kb; (2) are active at 37 °C; and (3) the DNA is fragmented with blunt ends, since blunt fragmentation avoids single-strand overhangs that can lead to artefactual double-strand mutations during end repair[8]. This in silico screen identified Hpy166II (recognition sequence: 5′-GTN/NAC-3′) as the optimal restriction enzyme, with a prediction of 37% of the genome mass fragmenting between 1 and 4 kb. The percentage of the genome fragmented to sizes between 1 and 4 kb was then empirically measured by fragmenting 1 µg of genomic DNA followed by quantification on a Genomic DNA ScreenTape assay (Agilent). Hpy166II fragments 41% of the genome to within the target size range. Note that, although Hpy166II is blocked by methylated CpG when present on both sides of the recognition sequence (New England BioLabs), this will occur only with the larger recognition sequence 5′-C*GTN/NAC*G-3′ (the asterisks signify methylation of the preceding cytosine); excluding all of these potential bimethylated sites increases the in silico predicted percentage of the genome fragmented by Hpy166II to within the target size range by 0.2%, and 99.97% of genomic bases within the original target size range remain when excluding these as potential fragmentation sites.

For the mitochondrial genome, Hpy166II captures 3 fragments in our target 1–4 kb size range, at the following coordinates (CHM13 v.1.0): (1) 3068–5116 (2,048 bp); (2) 7581–9439 (1,858 bp); and (3) 10441–11831 (1,390 bp). These fragments encompass 32% of the mitochondrial genome.

HiDEF-seq library preparation. Input DNA amounts of 500–3,000 ng (as measured using the Qubit 1× dsDNA HS Assay (Qubit)) were used per library, depending on available DNA. With high-quality DNA, input amounts of 500 ng provide sufficient HiDEF-seq library yield for approximately one full (non-multiplexed) Pacific Biosciences (PacBio) Sequel II instrument sequencing run, and lower input amounts are feasible for filling a fraction of a sequencing run. We have successfully made HiDEF-seq libraries with as low as 200 ng input DNA, producing sufficient yield for 40% of a sequencing run. For fragmented DNA samples, more than 1,500 ng of input DNA is generally required. Generally, for samples other than sperm and tissues from young children that have low mutation burdens, one quarter of a sequencing run is sufficient for mutation burden and pattern analysis. Input DNA $A_{260}/A_{280} > 1.8$ and $A_{260}/A_{230} > 2.0$ absorption ratios were confirmed on the NanoDrop before library preparation according to the Pacific Biosciences DNA preparation guidelines; we found that this quality control is important for sequencing yield for post-mortem tissues, but is not strictly necessary for other sample types.

As some DNA fragments are <1 kb after restriction enzyme fragmentation, these small fragments need to be removed during library preparation. We found that effective removal of <1 kb DNA fragments with high-yield recovery of larger DNA fragments requires three size selections with a 75% dilution of AMPure PB beads (Pacific Biosciences) during library preparation. We also found that efficient size selection critically depends on a DNA concentration of <10 ng µl^{-1} in the input sample. Accordingly, before beginning library preparation, a sufficient volume of AMPure PB Beads was diluted with Elution Buffer

(Pacific Biosciences) to a final 75% AMPure PB bead volume/total volume solution to be used for all subsequent bead purifications and size selections. Below, 'diluted AMPure beads' refers to these diluted beads.

Input DNA was fragmented with 0.14 U μl^{-1} of Hpy166II restriction enzyme (NEB) in a 70 µl reaction with 1× CutSmart buffer (NEB) for 20 min at 37 °C. The fragmentation reaction was scaled to a higher volume if the input DNA was too dilute to accommodate a 70 µl reaction.

Fibroblast samples from lentivirus transduction experiments and fibroblasts frozen in sperm-freezing medium that were extracted using the Puregene method comprised around 30–70% residual RNA (based on a comparison of Qubit and NanoDrop quantification), which was not fully removed by the DNA extraction's RNase A digestion. For these samples, we added 0.5 µl of 100 mg ml^{-1} RNase A (Qiagen) after completion of the fragmentation reaction and incubated at room temperature for 1 min.

Next, the fragmentation reaction was diluted with NFW to a DNA concentration of 10 ng μl^{-1} (or not diluted if DNA is already <10 ng μl^{-1}) based on the Qubit concentration of the DNA measured before the fragmentation reaction. For the first bead purification/size selection, a ratio of 0.8× diluted AMPure beads volume to sample volume was used, with two 80% ethanol washes, and the DNA was eluted with 22 µl of 10 mM Tris pH 8. The DNA concentration was measured again with Qubit.

Nick ligation was then performed (except for the initial version of HiDEF-seq without nick ligation; Supplementary Note 1) in a 30 µl reaction with 3 µl of 10× rCutSmart Buffer (NEB), 1.56 µl of 500 µM β-nicotinamide adenine dinucleotide (NAD^+) (NEB) and 15 U of *Escherichia coli* DNA ligase (NEB). The nick ligation reaction was incubated at 16 °C for 30 min with the heated lid turned off.

The DNA was then diluted with 10 mM Tris pH 8 to 10 ng μl^{-1} (or not diluted if DNA is already <10 ng μl^{-1}) based on the Qubit concentration measured after the post-fragmentation reaction bead purification. For the second bead purification/size selection, a ratio of 0.75× diluted AMPure bead volume to sample volume was used, with two 80% ethanol washes, and the DNA was eluted with 22 µl of 10 mM Tris pH 8. The DNA concentration was measured with Qubit.

The DNA was then treated as described in ref. 8 in 30 µl volume reactions with 21 µl of input DNA, 1.5 µl NFW, 3 µl 10× NEBuffer 4 (NEB), 3 µl of 1 mM each dATP/ddCTP/ddGTP/ddTTP (dATP/ddBTP) (dATP, Thermo Fisher Scientific, R0141; ddCTP/ddGTP/ddTTP, Jena Bioscience NU-1019S) and 7.5 U of Klenow fragment 3′→5′ exo- (NEB), except for input DNA that is degraded (such as post-mortem kidney and liver) for which the reaction was performed without dATP. The reaction was incubated at 37 °C for 30 min. Next, a third bead-purification/size-selection step was performed: the reaction volume was diluted with 10 mM Tris pH 8 to 10 ng μl^{-1} DNA (or not diluted if DNA is already <10 ng μl^{-1}) based on the Qubit concentration measured before the Klenow reaction, followed by a ratio of 0.75× diluted AMPure bead volume to sample volume, with two 80% ethanol washes, and elution of DNA with 22 µl of 10 mM Tris pH 8. The eluted DNA was then adjusted to a total of 30 µl with 3 µl of 10× NEBuffer 4 and NFW before proceeding to adapter ligation.

Ligation of hairpin adapters was performed using reagents from the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) by combining 30 µl of Klenow-treated DNA, 2.5 µl Barcoded Overhang Adapter (Pacific Biosciences), 15 µl Ligation Mix, 0.5 µl Ligation Additive and 0.5 µl of Ligation Enhancer. For samples for which the preceding Klenow reaction was performed without dATP (that is, non-A tailed libraries), 2.5 µl of 17 µM annealed blunt adapters were used instead (their sequences and preparation are described below). The adapter ligation reaction was incubated at 20 °C for 60 min with the heated lid turned off. Immediately after the adapter ligation, nuclease treatment was performed using the SMRTbell Enzyme Clean Up Kit 1.0 (Pacific Biosciences) to remove any non-circularized DNA containing nicks and/or without hairpin adapters: 2 µl Enzyme A, 0.5 µl Enzyme B, 0.5 µl Enzyme C and 1 µl Enzyme D were combined, and this 4 µl enzyme mix was added to the ligation reaction and incubated at 37 °C for 60 min. After the nuclease treatment, the samples were purified with a ratio of 1.2× diluted AMPure bead volume to sample volume and eluted with 24 µl of 10 mM Tris pH 8.

After the post-nuclease treatment AMPure bead purification, non-A tailed HiDEF-seq libraries underwent an additional 1.1× diluted AMPure bead purification to remove residual adapter dimers.

Final library concentration and size distribution were measured using the Qubit and High Sensitivity D5000 ScreenTape (Agilent). The final library fragment size distribution should contain <5% of DNA mass <1 kb and <5% of DNA mass >4 kb (percentages calculated using the ScreenTape analysis software's manual region analysis '% of Total' field). The final mass yield of A-tailed libraries should be around 6–10% of the input genomic DNA mass, and approximately half of that for non-A tailed libraries. The libraries were stored in 0.5 ml DNA LoBind tubes at −20 °C.

On ScreenTape, some non-A tailed HiDEF-seq libraries had a low level of residual adapter dimers, which was removed with a final 1.3× diluted AMPure bead purification after multiplexing the libraries from the same run (see multiplexing details in the 'HiDEF-seq library sequencing' section).

Sequences and preparation of blunt adapters used for HiDEF-seq without A-tailing. Adapters (sequences below) were ordered as HPLC-purified oligonucleotides from Integrated DNA Technologies. Each adapter was reconstituted to 100 µM concentration with NFW. Annealing was then performed for each adapter at a concentration of 17 µM in a 30 µl volume containing 10 mM Tris pH 8 and 50 mM NaCl, by incubating at 95 °C for 3 min and cooling at room temperature for 30 min. Additional barcoded adapters can be designed by replacing the below barcodes with alternative sequences. bcAd1001: /5′Phos/ACG CACTCTGATATGTGATCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTT GAGAGAGATCACATATCAGAGTGCGT (barcode = CACATATCAGAGT GCG); bcAd1002: /5′Phos/ACTCACAGTCTGTGTGTATCTCTCTCTTTTC CTCCTCCTCCGTTGTTGTTGTTGAGAGAGATACACACAGACTGTGAGT (barcode = ACACACAGACTGTGAG); bcAd1003: /5′Phos/ACTCTCACGA GATGTGTATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAG AGATACACATCTCGTGAGAGT (barcode = ACACATCTCGTGAGAG); bcAd1008: /5′Phos/ACGCAGCGCTCGACTGTATCTCTCTCTTTTC CTCCTCCTCCGTTGTTGTTGTTGAGAGAGATACAGTCGAGCGCTGCGT (barcode = ACAGTCGAGCGCTGCG).

Modified HiDEF-seq library preparation trials to remove ssDNA T>A artefacts. Below are details of trials to remove ssDNA T>A artefacts arising from ssDNA nicks that remain after nick ligation. The final protocol that completely removes these artefacts (HiDEF-seq without A-tailing) is described in the main 'HiDEF-seq library preparation' section.

Polynucleotide kinase. The standard HiDEF-seq protocol was followed with the exception that, before nick ligation, the DNA was treated in a 30 µl reaction containing 12 U T4 polynucleotide kinase (NEB), 1 mM ATP (NEB), 4 mM DTT (Promega) and 1× CutSmart buffer (NEB) at 37 °C for 1 h. The sample then proceeded into the nick ligation reaction in a higher reaction volume of 35 µl, with reaction components scaled proportionally to the higher volume and a final 1× CutSmart buffer concentration.

Alternative A-tailing polymerases. The standard HiDEF-seq protocol was followed with the exception of replacing Klenow fragment 3′→5′ exo- polymerase with one of the following: 9.6 U Bst large fragment (NEB), 9.6 U Bst 2.0 (NEB), 9.6 U Bst 3.0 (NEB) or 9 U Isopol SD+ (ArcticZymes). The reaction temperatures and times for these polymerases were as follows: (1) Bst large fragment and Bst 2.0: 30 min at 45 °C; (2) Bst 3.0: 30 min or 150 min at 45 °C, or 210 min at 37 °C; and (3) Isopol SD+: either 30 min or 210 min at 37 °C.

Pyrophosphatase. The standard HiDEF-seq protocol was followed with the exception of adding 0.15 U of *E. coli* inorganic pyrophosphatase (NEB).

Klenow reaction without dATP or without dATP/ddBTP. The standard HiDEF-seq protocol was followed with the exception that the Klenow reaction was performed without dATP or without dATP/ddBTP.

No Klenow reaction. The standard HiDEF-seq protocol was followed, except that, after the post-nick ligation bead purification, the DNA was diluted to 30 µl in a final 1× NEBuffer 4 concentration and taken directly to adapter ligation using blunt adapters. After the post-nuclease treatment bead purification, an additional size-selection step was performed with 0.75× diluted AMPure beads as this would ordinarily have occurred after the Klenow reaction. Note that this protocol produces a CCT>CGT ssDNA artefact that does not occur when the Klenow reaction is performed without dATP or ddBTP, indicating that Klenow polymerase removes this artefact likely through a pyrophosphorolysis mechanism (Extended Data Fig. 5d and Supplementary Table 3).

HiDEF-seq library preparation with uracil DNA glycosylase/endonuclease VIII treatment. Libraries were prepared according to the above HiDEF-seq library protocol with A-tailing, except that 3 µl of a mixture of uracil DNA glycosylase/endonuclease VIII (NEB USER enzyme, M5505) was added to the nuclease treatment step.

HiDEF-seq library preparation with multi-glycosylase/endonuclease treatment. Libraries were prepared according to the above HiDEF-seq library protocol without A-tailing, except that, after the bead purification/size selection that occurs after the Klenow ddBTP reaction, an additional enzyme treatment step was performed. This enzyme treatment occurred in a total volume of 60 µl in a final 1× ThermoPol Buffer (NEB) at 37 °C for 30 min, with the following enzymes: (1) 10 U endonuclease IV (NEB); (2) 8 U formamidopyrimidine DNA glycosylase (Fpg) (NEB); (3) 10 U T4 pyrimidine dimer glycosylase (NEB); (4) 2 µl of a mixture of uracil DNA glycosylase/endonuclease VIII (NEB USER enzyme); (5) 10 U endonuclease VIII (NEB); (6) 10 U human alkyl adenine DNA glycosylase (hAAG) (NEB); and (7) 5 U human single-stranded selective monofunctional uracil DNA glycosylase (hSMUG1) (NEB). This reaction was cleaned up with a ratio of 1.2× diluted AMPure bead volume to sample volume, with two 80% ethanol washes, and elution of DNA with 22 µl of 10 mM Tris pH 8. The eluted DNA was then adjusted to a total of 30 µl with 3 µl of 10× NEBuffer 4 and NFW before proceeding to adapter ligation according to the standard HiDEF-seq protocol.

HiDEF-seq large fragment library preparation. Large-fragment-size libraries (range, 1–10 kb; median, 4.1 kb) were prepared according to the above HiDEF-seq library protocol, except (1) fragmentation was performed with 30 U PvuII-HF enzyme (NEB) instead of Hpy166II; (2) post-nick ligation and post-A-tailing cleanups were performed with 1.8× undiluted AMPure PB beads, and DNA was not diluted to <10 ng µl^{-1} (since size selection is not being performed); and (3) final post-nuclease treatment bead purification was performed with 1× undiluted AMPure PB beads.

HiDEF-seq library preparation with random fragmentation. Libraries were prepared according to the above HiDEF-seq library protocol without A-tailing (that is, Klenow reaction without dATP and using blunt adapters), except that (1) a higher amount of input DNA was used (4 µg per sample); (2) instead of restriction enzyme fragmentation, DNA was acoustically fragmented in miniTUBE Clear tubes (2 µg per tube, that is, 2 × 2 µg aliquots per sample), with each 2 µg DNA aliquot diluted to 200 µl in a final buffer of 10 mM Tris pH 8 and 50 mM NaCl, on an ME220 instrument (Covaris) using the following settings: temperature, 7 °C; treatment time, 900 s; peak incident factor, 8 W; duty factor, 20%; and cycles/burst, 1,000; (3) each 2 µg fragmented DNA aliquot was blunted in a 200 µl reaction containing 0.5 U µl^{-1} nuclease P1 (NEB) and 1× NEBuffer r1.1 (NEB) at 37 °C for 30 min, after which the reaction was stopped by adding 8 µl of 0.5 M EDTA and 2 µl of 1% SDS; (4) after the Nuclease P1 reaction, the protocol continued with the 0.8× diluted AMPure bead purification as is usually performed after restriction enzyme fragmentation, and the two aliquots of each sample were combined at the elution stage for a final elution volume of 22 µl; (5) before nick ligation, the DNA was treated with 0.4 U µl^{-1} T4 polynucleotide kinase (NEB), 1 mM ATP and 4 mM DTT in a 30 µl volume of 1× rCutSmart buffer (NEB) at 37 °C for 1 h; (6) nick ligation was performed immediately after by adding the required reagents to the T4 polynucleotide kinase reaction to a final volume of 35 µl; (7) the bead-purification step after the Klenow reaction was performed with a 1.2× ratio of diluted AMPure bead volume to sample volume, instead of a ratio of 0.75×; (9) after nuclease treatment, libraries underwent a 1.2× diluted AMPure bead purification, then libraries for the same sequencing run were pooled, and a final 1.0× diluted AMPure bead purification was performed to remove residual adapter dimers.

HiDEF-seq library sequencing. Libraries sequenced on the same sequencing run were multiplexed together based on the final library Qubit quantification to achieve at least 50 ng of total library in no more than 15 µl volume. When necessary, the concentration of individual or pooled libraries can be increased by room temperature centrifugal vacuum concentration (Eppendorf Vacufuge) and pausing periodically (approximately every 2 min) to avoid increases in temperature, or using AMPure PB bead purification.

Sequencing was performed on Pacific Biosciences Sequel II or Sequel IIe systems with 8M SMRT Cells by the Icahn School of Medicine at Mount Sinai Genomics Core Facility and the New York University Grossman School of Medicine Genome Technology Center. Sequencing parameters were as follows: Sequel II Binding Kit 2.0 (Pacific Biosciences), Sequel II Sequencing Kit 2.0 (Pacific Biosciences), Sequencing primer v4 (Pacific Biosciences), 1 h binding time, diffusion loading, loading concentrations between 125 and 160 pM (lower concentrations were used for blood than for tissues) for standard size libraries (Hpy166II libraries) or 80 pM for large-fragment libraries (PvuII libraries), 2 h pre-extension, and 30 h movies.

Germline and tumour sequencing data processing

The HiDEF-seq computational pipeline can filter germline variants using either standard short-read or long-read genome sequencing of the same individual (Extended Data Fig. 3k,l).

Illumina germline and tumour sequencing data processing. Reads were aligned to the CHM13 v.1.0 reference genome[58] using BWA-MEM (v.0.7.17)[59] with the standard settings, followed by marking of optical duplicates and sorting using the Picard Toolkit v3.1.0 (Broad Institute). Variants were called from the aligned reads with two different variant callers: (1) Genome Analysis Toolkit (GATK)[60] v.4.1.9.0 using the HaplotypeCaller tool with the parameters '-ERC GVCF -G StandardAnnotation -G StandardHCAnnotation -G AS_StandardAnnotation' followed by the GenotypeGVCFs tool with the default parameters; (2) DeepVariant[61] v.1.2.0 with the parameter: '--model_type=WGS'. Both GATK and DeepVariant variant calls were used during the subsequent analysis.

Pacific Biosciences germline sequencing data processing. Circular consensus sequences were derived from raw subreads (a subread is one sequencing pass of a single strand of a DNA molecule) using pbccs v.5.0.0 (ccs, Pacific Biosciences) with the default parameters, and consensus sequences were filtered to retain only high-quality 'HiFi' reads, that is, reads with predicted consensus sequence accuracy 'rq' tag ≥ 0.99 (rq is calculated by ccs as the average of the per base consensus qualities of the read). These reads were then aligned to the CHM13 v.1.0 reference genome with pbmm2 v.1.4.0 (Pacific Biosciences)

with the parameters '--preset CCS --sort'. Variants were called from the aligned reads with DeepVariant[61] v.1.2.0 with the parameter '--model_type=PACBIO'.

HiDEF-seq primary data processing

HiDEF-seq data first undergoes a two-part primary data processing pipeline that transforms the raw data into a format suitable for subsequent analysis. Primary data processing also produces quality-control plots generated by custom scripts and by SMRT Link (Pacific Biosciences) software (for example, distributions of polymerase read lengths and number of passes). Note that, for simplicity, we use the term 'call' to refer to both dsDNA mutations and ssDNA mismatch and damage events. The pipeline analyses calls in sequencing reads that are single-base mismatches relative to the reference genome (that is, not insertions and deletions).

The first part of the primary data-processing pipeline uses a combination of bash and awk scripts to process raw subread sequencing data (a subread is one sequencing pass of a single strand of a DNA molecule) into a strand-specific aligned BAM format[62] with additional tags needed for call analysis[62]. The steps of this first part of data processing are as follows:

(1) Subreads for which an adapter was not detected on both ends of the molecule ('cx' tag not equal to 3) are removed.
(2) Consensus sequences are created separately for each strand of the DNA molecule (that is, forward and reverse strand separately) using pbccs v.6.2.0 (Pacific Biosciences) with the parameters: --by-strand, --min-rq 0.99 (minimum predicted consensus sequence accuracy > Q20 (Phred quality score) to remove low-quality consensus sequences) and --top-passes 0 (unlimited number of full-length subreads used per consensus).
(3) Demultiplexing of samples according to adapter barcodes using lima v.2.5.0 (Pacific Biosciences) with the parameters: --ccs --same --split-named --min-score 80 --min-end-score 50 --min-ref-span 0.75 --min-scoring-regions 2.
(4) Filter to remove any DNA molecules (also known as zero-mode waveguides, which are sequencing wells containing a single DNA molecule) that did not successfully produce both one forward- and one reverse-strand consensus sequence.
(5) Align forward- and reverse-strand consensus sequences to the CHM13 v.1.0 reference genome[58] using pbmm2 v.1.7.0 (Pacific Biosciences), an aligner based on minimap2[63], with the parameters: --preset CCS. We use the telomere-to-telomere CHM13 human reference genome, which was itself constructed using long reads, to reduce genome alignment artefacts[58]. Note that the CHM13 v.1.0 reference genome contains only nuclear chromosomes 1–22, chromosome X and the mitochondrial genome—but not chromosome Y, which is therefore not part of the analyses.
(6) Filter to retain only DNA molecules that produce only one forward strand primary not-supplementary alignment and one reverse-strand primary not-supplementary alignment, where the forward and reverse alignments overlap (reciprocally) in the genome by at least 90%.
(7) Sort alignments by reference position.
(8) Add five tags, detailed below, to each alignment in the final BAM file, with each tag containing a comma-separated array with a length corresponding to the number of single-base mismatches in the alignment (relative to the reference genome) per the alignment CIGAR string:
   - qp: positions of bases in the read sequence (query) that are mismatches relative to the reference genome; 1-based coordinates with the left-most base in the alignment record's SEQ column = 1;
   - qn: sequences of bases in the read (query) that are mismatches relative to the reference genome (base sequences are according to the forward genomic strand, that is, they are taken from the SEQ column of the SAM alignment record);
   - qq: qualities of bases in the read that are mismatches relative to the reference genome (taken from the QUAL column of the SAM alignment record);
   - rp: positions in reference genome coordinates of read bases that are mismatches relative to the reference genome;
   - rn: sequences of the reference genome at positions of read bases that are mismatches relative to the reference genome.

The second part of the primary data processing pipeline is an R[64] script (R v.4.1.2, requiring the packages Rsamtools[65], GenomicAlignments[66], GenomicRanges[66], vcfR[67], plyr[68], configr[69], qs[70]) that further processes and annotates the aligned BAM file into an R data file as follows:

(1) Load the aligned BAM file into R, including the custom tags that annotated the positions of base mismatches relative to the reference genome.
(2) Annotate calls (bases mismatched relative to the reference genome) for which the reference genome base is '*N*', to exclude these from subsequent analysis.
(3) Annotate the positions of indels in each alignment, based on the alignment CIGAR string.
(4) Annotate each call if it was present in any of the VCF variant call files of the corresponding individual's germline sequencing, along with details of the VCF variant annotation.
(5) Save positions of indels from the VCF variant call files of the corresponding individual's germline sequencing.
(6) Transform the dataset so that forward- and reverse-strand consensus reads and ssDNA and dsDNA calls (and tag information) from the same DNA molecule are linked to each other as dsDNA molecules.
(7) Save the final R dataset to a file.

HiDEF-seq call filtering

The call-filtering pipeline implements a series of filters that were optimized to maximize the number of true calls while minimizing the number of sequenced bases and regions of the genome that are filtered out. During the development of the pipeline, filters and filter parameters were iteratively optimized using low-mutation-rate samples (that is, tissues from infants and sperm) to identify patterns that are common to false positives. These false positives were apparent as clusters of mutations in low-quality regions of the genome and as regions with low-quality alignment of sequencing reads. For example, when a metric of low-quality genome regions was found to correlate with clusters of low-quality calls, this metric was added as a filter, and its threshold was iteratively tuned to maximally remove false positives while minimizing the number of sequenced bases and genomic regions that are filtered.

Additional optimization of filter thresholds was performed using sperm samples that have a known low mutation burden. Specifically, we plotted the dsDNA and ssDNA burdens with a range of thresholds for three key filters: (1) minimum predicted consensus accuracy (0.99 to 0.999); (2) minimum number of passes per strand (5 to 20); and (3) minimum fraction of subreads (passes) detecting the mutation (0.5 to 0.8) (Extended Data Fig. 3c–j). We examined these plots for threshold settings above which burden estimates are stable. Since burdens were corrected for sensitivity (based on total interrogated bases and detection of known germline variants; see the 'HiDEF-seq calculation of call burdens' section), a decrease in burden estimates with increasing threshold settings indicates removal of sequencing artefacts. These plots showed that sperm dsDNA mutation burden estimates were stable even down to the lowest (most lenient) thresholds (Extended Data Fig. 3d,e,g). By contrast, ssDNA burdens required higher threshold settings before burden estimates stabilized (Extended Data Fig. 3i,j). Individually increasing the thresholds of each of the above three filters stabilized ssDNA burden estimates at approximately 20%, 15% and 10% lower levels, respectively, compared to the least stringent settings, and applying all three filters together with these higher thresholds

reduced the ssDNA burden estimate by approximately 25% (that is, the three filters are not independent). Specific thresholds used for dsDNA and ssDNA mismatch filtering are detailed in the below sections detailing each filter.

The call-filtering pipeline uses the following R packages: GenomicAlignments (v.1.30.0)[66], GenomicRanges (v.1.46.1)[66], vcfR (v.1.12.0)[67], Rsamtools (v.2.10.0)[65], plyr (v.1.8.6)[68], configr (v.0.3.5)[69], MutationalPatterns (v.3.4.1)[71], magrittr (v.2.0.2)[72], readr (v.2.1.2)[73], dplyr (v.1.0.8)[74], plyranges (v.1.14.0)[75], stringr (v.1.4.0)[76], digest (v.0.6.29)[77], rtracklayer (v.1.54.0)[78], qs (v.0.25.2)[70]; and the following software tools: bcftools (v.1.14)[79], samtools[79], wigToBigWig (v.2.8)[80], wiggletools (v.1.2.11)[81], pbmm2 (v.1.7.0; Pacific Biosciences), zmwfilter (v.1.2.0; Pacific Biosciences), SeqKit (v.2.1.0)[82] and KMC (v.3.1.1)[83].

Additional filters used in the pipeline were created using REAPR (v.1.0.18)[84]. REAPR was originally designed to identify regions with errors in reference genome assemblies, but we found that it calculates metrics that are useful for identifying regions of the genome prone to generating false-positive and false-negative variant calls in Illumina (short-read) sequencing data. First, Illumina whole-genome sequencing reads from a sperm sample were aligned to CHM13 v.1.0 using SMALT (v.0.7.6)[85] with the parameters '-r 0 -x -y 0.5' and a CHM13 v.1.0 index created with SMALT using parameters '-k 13 -s 2'. Next, reads were sorted and duplicates were marked. The REAPR perfectfrombam command was then run on the resulting BAM file using the parameters 'min insert=266, max insert=998, repetitive max qual=3, perfect min qual=4, and perfect min alignment score=151' (min and max insert size are the 1 and 99 percentiles of insert sizes calculated from the sequencing data using the Picard Toolkit CollectInsertSizeMetrics tool). REAPR metrics for each base of the genome were obtained from the output stats.per_base file and a bigwig[86] annotation file was created for each metric.

The mutation analysis filters were applied serially as described below. Unless otherwise specified, the filters were applied to both ssDNA and dsDNA calls. Note that the computational pipeline has the capability to implement additional filters not listed here, as specified in the pipeline configuration documentation available online.

Filters based on DNA molecule quality and alignment metrics. Retain only DNA molecules that meet all of the below criteria:

(1) ccs predicted consensus accuracy ≥ 0.99 in both forward and reverse strand (that is, rq tag of ccs ≥ 0.99) for dsDNA calls, and ≥Q30 (that is, rq ≥ 0.999) for ssDNA calls.
(2) Minimum of 5 (for dsDNA calls) and 20 (for ssDNA calls) sequencing passes for each of the forward and reverse strands (using the 'ec' BAM file tag, which is computed by ccs as the average subread coverage across all consensus calling windows).
(3) Both forward and reverse strands have mapping quality ≥ 60.
(4) Maximum difference in number of ssDNA calls between the forward and reverse strands of 5, before germline variant filtering. This removes artefacts from rare chimeric molecules and residual low-quality molecules.
(5) Average of the number of indels relative to the human reference genome in the forward and reverse strands of ≤20, before germline variant filtering. This removes low-quality molecules with many indels.
(6) Average of the number of soft-clipped bases in the forward and reverse strands of ≤30. This removes low-quality molecules and molecules that align to complex regions of the genome with long stretches of mismatched bases.

Filters based on germline sequencing variant calls.

(1) Filter out calls that were also identified in any of the individual's germline sequencing VCF files with read depth ≥3, allele quality (QUAL column in VCF) ≥ 3, genotype quality (GQ tag in VCF) ≥ 3, and variant allele fraction ≥ 0.05.
(2) Filter out DNA molecules with >8 dsDNA calls remaining after VCF germline filtering. This removes molecules with misalignment to complex regions of the genome leading to many clustered calls and regions of the genome for which Illumina short reads are not effective in identifying and filtering out germline variants.

For tumour analysis, variant calls were used in this step from both germline blood sequencing and standard fidelity (Illumina) tumour sequencing to focus the analysis on low-level mosaic calls.

Filters based on genomic regions. Filters that remove the entire DNA molecule if it meets any of the following criteria:

(1) For analyses using either Illumina or PacBio germline sequencing data: (i) segmental duplication regions: any overlap with the DNA molecule's forward or reverse consensus sequence alignments. This annotation was obtained from the file chm13.draft_v1.0_plus38Y.SDs.bed created by the Telomere-to-Telomere consortium[87]. However, for analysis of mitochondrial mutations, this region filter is not used because it contains the region chrM:10000–14910 due to a similar nuclear genome sequence on chromosome 5, which would cause unnecessary filtering of reads aligning to this region of the mitochondrial genome. There is negligible risk of nuclear genome sequences falsely aligning to this mitochondrial region since we obtain long reads, we require high mapping quality, and we exclude reads with many mismatches—and these mitochondrial and nuclear genome regions have only 94% identity. (ii) Satellite sequence regions: ≥20% of the DNA molecule's forward- and reverse-strand consensus alignments (average for the two strands) overlaps the region. The satellite sequence region annotation was created for CHM13 v.1.0 using RepeatMasker (v.4.1.1)[88] with the parameters '-pa 4 -e rmblast -species human -html -gff -nolow', followed by extraction of 'Satellite' regions.
(2) Only for analyses that use Illumina germline sequencing data, because short-read data is more prone to missing true germline variants in these regions: (i) telomere regions: any overlap with the DNA molecule's forward or reverse consensus sequence alignments. This annotation was obtained from the file chm13.draft_v1.0.telomere created by the Telomere-to-Telomere consortium[58]. (ii) 50-mer mappability score: ≥30% of the DNA molecule's forward- and reverse-strand consensus alignments (average for the two strands) has a mappability score < 0.4. This annotation was created for CHM13 v.1.0 using Umap (v.1.2.0)[89]. This annotation calculates the mappability for every base in the genome. (iii) The fraction of Illumina short reads aligning to the region that are orphaned reads (that is, the read's mate is either unmapped or mapped to a different chromosome), averaged across the genome in 20 bp non-overlapping bins, is ≥0.15 for ≥20% of the DNA molecule's forward- and reverse-strand consensus alignments (average for the two strands). The fraction of orphaned reads metric used in this filter is the average of the orphan_cov and orphan_cov_r REAPR metrics, which are the fraction of forward- and reverse-strand reads that are orphaned, respectively.

Filters that remove only the portions of the DNA molecule that overlap any of the following regions, while the remaining bases of the DNA molecule are still included in analysis:

(1) Regions of the reference genome whose sequence is '*N*'.
(2) For analyses using either Illumina or PacBio germline sequencing data: (i) satellite sequence regions: any base that overlaps one of these regions. (ii) Bases with gnomAD (v.3.1.2)[90] single-nucleotide variants with 'PASS' flag and population allele frequency > 0.1%, lifted over from the hg38 to the CHM13 v.1.0 reference genome using the liftOver tool[80]. This filter removes 27,476,828 genome bases from the analysis. It is required to remove residual germline variants that were not detected in the germline sequencing of the

individual, and it reduces the risk of false-positive mosaic event calls due to very low level contamination that may occur between samples of different individuals[8].

(3) Only for analyses that use Illumina germline sequencing data, because short-read data are more prone to missing true germline variants in these regions: (i) 100-mer mappability score: any base with a mappability score < 0.95, with mappability scores averaged across the genome in 20 bp non-overlapping bins (binning smoothes the mappability score signal). The primary mappability scores were calculated as described for the above 50-mer mappability score. (ii) The fraction of Illumina short reads aligning to the region that are properly paired (that is, aligned in the correct orientation and within the expected distance based on insert size distribution), averaged across the genome in 20 bp non-overlapping bins, is <0.7. The fraction of properly paired reads metric used in this filter is the average of the prop_cov and prop_cov_r REAPR metrics, which are the fraction of forward-strand and reverse-strand reads that are properly paired, respectively. (iii) The fraction of Illumina short reads aligning to the region that are orphaned reads (that is, the read's mate is either unmapped or mapped to a different chromosome), averaged across the genome in 20 bp non-overlapping bins, is ≥0.2. The fraction of orphaned reads metric used in this filter is the average of the orphan_cov and orphan_cov_r REAPR metrics, which are the fraction of forward- and reverse-strand reads that are orphaned, respectively. (iv) The number of Illumina short reads aligning to the region to either the forward or the reverse strand and that are soft-clipped at the left end or the right end (that is, the sum of the REAPR clip_fl, clip_fr, clip_rl, clip_rr metrics), divided by [4 × number of mapped reads/100,000,000], averaged across the genome in 200 bp non-overlapping bins, is ≥0.09. (v) The number of Illumina short reads with mapping quality 0 aligning to the region, divided by [4 × number of mapped reads/100,000,000], averaged across the genome in 20 bp non-overlapping bins, is ≥0.1. Note that this general filtering annotation was calculated using Illumina whole-genome sequencing data of one representative sample.

Base quality filter. Filter out dsDNA calls whose consensus sequence base quality is <93 (from QUAL column in BAM file) in either the forward- or reverse-strand consensus, and filter ssDNA calls whose base quality is <93 in the strand containing the call.

Filter based on location within the read. Filter out calls that are ≤10 bases from the ends of the consensus sequence alignment (alignment span excludes soft-clipped bases). For ssDNA calls, this filter is applied to the strand containing the call and, for dsDNA calls, this filter is applied to both the forward- and reverse-strand consensus sequence alignments. Although this only negligibly alters call burdens (Extended Data Fig. 3h), it removes rare alignment artefacts.

Filter based on location near germline indels. Regions near germline indels are prone to alignment artefacts that can lead to false-positive calls. This filter removes calls located near an indel within a distance less than or equal to twice the length of the indel or less than or equal to 15 bases of the indel (whichever range is larger), using indels called in any of the germline sequencing data of the individual (that is, both GATK and DeepVariant indel calls when using Illumina germline sequencing data, and only DeepVariant indel calls when using PacBio germline sequencing data). For GATK indel calls, only indels with read depth ≥ 5, QUAL ≥ 10, genotype quality ≥ 5 and variant allele fraction ≥ 0.2 were used in this filtering. For DeepVariant indel calls, only indels with read depth ≥ 3, QUAL ≥ 3, genotype quality ≥ 3 and variant allele fraction ≥ 0.1 were used in this filtering.

Filter based on location near consensus sequence indels. Regions near HiDEF-seq consensus sequence indels are prone to alignment artefacts that can lead to false-positive calls. This filter removes calls located near a consensus sequence indel within a distance less than or equal to twice the length of the indel or less than or equal to 15 bases of the indel (whichever range is larger). For dsDNA calls, the call must pass this filter on both forward and reverse consensus strands. For ssDNA calls, this filter applies only to the strand containing the call.

Filters based on germline sequencing read depth and variant allele fraction.

(1) Filter out calls in locations where the germline sequencing data has <15 total reads coverage, as these low-coverage germline sequencing regions will be prone to false-negative germline variant calls that would then lead to false-positive HiDEF-seq calls.

(2) Filter out calls detected with variant allele fraction > 0.05 or read depth > 3 in the germline sequencing data to remove variants that were not called by the previous germline variant callers (due to low variant allele fraction or due to different local haplotype assembly in GATK/DeepVariant that calls variants in a different nearby location than the bwa alignment of the consensus molecule sequence). This filter is less stringent than a recent somatic mutation analysis method[8], but may still remove a small number of very early developmental mosaic variants shared between HiDEF-seq data and the individual's germline sequencing.

The above two filters use the samtools mpileup command to determine total read depth and variant allele fraction, using the parameters '-I -A -B -Q 11 --ff 1024 -d 10000 -a "INFO/AD"' for Illumina germline sequencing data and the parameters '-I -B -Q5 --ff 2048 --max-BQ 50 -F0.1 -o25 -e1 --delta-BQ 10 -M399999 -d 10000 -a "INFO/AD"' for PacBio germline sequencing data.

For tumour analysis, this filter step used both germline blood sequencing and standard fidelity (Illumina) tumour sequencing to focus the analysis on low-level mosaic calls.

Filters based on fraction of subreads (passes) detecting the call and fraction of subreads overlapping the call. We filter out calls detected in <50% (for dsDNA calls) and <60% (for ssDNA calls) of the subreads of the DNA molecule that detected the call. For dsDNA calls, this filter is applied to forward and reverse subreads separately, and the call must pass the filter in both strands. For ssDNA calls, this filter is applied only to subreads of the strand in which the call was detected.

This removes false-positive calls in the consensus sequence that are not well-supported by the subreads. The filter is implemented by first extracting the subreads of all of the DNA molecules containing calls from the raw subreads BAM file using the zmwfilter tool (Pacific Biosciences) and aligning them to the CHM13 v.1.0 reference genome with pbmm2 v.1.7.0 with the parameters '--preset SUBREAD --sort'. The bcftools mpileup command is then used with the parameters '-I -A -B -Q 0 -d 10000 -a "INFO/AD"' to calculate the fraction of subreads detecting the call (excluding subreads with the supplementary alignment SAM flag).

In rare DNA molecules, a large fraction of subreads is soft-clipped, leading to false-positive calls in the small fraction of remaining subreads aligned to the soft-clipped region. We therefore also filter out calls for which the percentage of subreads overlapping the call (regardless of whether they contain the call) out of the total subreads aligned to the genome is <50%, calculated separately for subreads of each strand for the molecule in which the call was made. This filter is applied to the strand containing the call for ssDNA calls, and to both strands for dsDNA calls (that is, a dsDNA call must pass this filter in both strands).

HiDEF-seq calculation of call burdens

After application of all of the above filters, DNA molecules are further filtered to retain only those with a maximum of one dsDNA call for dsDNA call burden calculations, and a maximum of one ssDNA call per

strand for ssDNA call burden calculations. This removes a small number of the remaining DNA molecules that contain multiple post-filtering calls that, after manual inspection, are due to residual regions of the genome prone to false positives.

The raw dsDNA mutation burden (that is, mutations per bp) of a sample is then calculated as the [number of dsDNA calls]/[number of interrogated dsDNA base pairs], and the raw ssDNA call burden (that is, calls per base) is calculated as the [number of forward strand calls + number of reverse strand calls]/[number of interrogated forward strand read bases + number of interrogated reverse strand read bases]. Note that we subsequently use the term 'interrogated bases' for simplicity, even though, for dsDNA mutation analysis, it refers to interrogated base pairs. The number of interrogated bases takes into account all of the relevant filters that were applied, both filters that remove entire DNA molecules and filters that remove only portions of DNA molecules. Specifically, the number of interrogated bases is the total number of bases of DNA molecules that passed all of the filters that remove full DNA molecules (described in the 'Filters based on DNA molecule quality and alignment metrics' and 'Filters based on genomic regions' (first part) sections), excluding the bases of those remaining DNA molecules removed by the following filters (described above) that remove only portions of DNA molecules: (1) 'Filters based on genomic regions' (second part); (2) 'Base quality filter'; (3) 'Filter based on location within the read'; (4) 'Filter based on location near germline indels'; (5) 'Filter based on location near consensus sequence indels'; and (6) the minimum germline sequencing total read coverage filter described in the 'Filters based on germline sequencing read depth and variant allele fraction' section.

We also calculated 'corrected' call burdens that correct both for: (1) differences in trinucleotide sequence context of the genome relative to interrogated bases; and (2) sensitivity of detection. These corrections were applied as follows:

First, we corrected raw call counts for the trinucleotide frequency distribution of the genome (specifically, the CHM13 v.1.0 sequences of chromosomes being analysed; that is, chromosomes 1–22 and X for nuclear genome analysis, and the mitochondrial sequence for mitochondrial genome analysis) relative to the trinucleotide frequency distribution of interrogated bases in sequencing reads. This correction for 'trinucleotide context opportunities' is necessary because interrogated bases may have a different distribution of trinucleotides compared to the genome due to restriction enzyme fragmentation and computational filters, and this may affect burden estimates[8]. Specifically, we first calculate the distribution of trinucleotides (the fraction of each trinucleotide out of all trinucleotides) across the genome. We then calculate the distribution of trinucleotides across interrogated bases of sequencing reads in the sample. Next, for each trinucleotide, we calculate the ratio of its fractional distribution in the full genome to its fractional distribution in the interrogated bases. The trinucleotide-corrected count of HiDEF-seq calls is then obtained by multiplying the raw call count for each trinucleotide context by that context's genome/interrogated bases trinucleotide ratio. For dsDNA calls, trinucleotide context corrections are performed using all possible 32 trinucleotide contexts where the middle base is a pyrimidine. For ssDNA calls, trinucleotide context corrections are performed using all 64 possible trinucleotides and using strand-specific trinucleotide sequences of calls, interrogated bases and the genome. The trinucleotide contexts of ssDNA calls reflect the original DNA molecule's ssDNA change—that is, for calls in strands aligning to the forward strand of the reference genome, the reverse complements of the call, interrogated read sequences and genome are used for trinucleotide context corrections, and vice versa for calls in strands aligning to the reverse strand. This is because the sequence data produced by the sequencer has the directionality of the sequencer-synthesized strand rather than the original (template) DNA molecule.

Second, we corrected call counts for sensitivity of detection separately for each sample using a set of high-quality, true-positive heterozygous germline (dsDNA) variants detected in the HiDEF-seq data of the sample. This specifically accounts for single-molecule sensitivity loss due to the 'Filters based on fraction of subreads (passes) detecting the call and fraction of subreads overlapping the call' that are applied to calls detected in the final interrogated bases (they are applied to each strand separately, and dsDNA calls must pass the filters in both strands). All of the other filters remove DNA molecules and bases from the final set of interrogated bases and therefore do not require sensitivity correction. To generate the true-positive set of heterozygous germline variants for each sample, we extracted all of the autosomal dsDNA calls detected in the final interrogated HiDEF-seq bases of the sample that were also called in all of the germline variant call sets of the individual with ≥50th percentile VCF QUAL score, ≥50th percentile VCF genotype quality, ≥50th percentile total read depth, and variant allele fraction between 30% and 70%. We retain only calls that meet these criteria across every one of the variant call sets of the individual and that are present in gnomAD v.3.1.2 with 'PASS' flag and population allele frequency > 0.1%. If more than 10,000 such true-positive germline calls are identified, a random subset of 10,000 calls is selected for the sensitivity calculation. We then extract subreads corresponding to the DNA molecules that detected these calls in the sample, realign them to the genome with pbmm2 v.1.7.0 with the '--preset SUBREAD --sort' settings and annotate the variants using the same process described in the 'Filters based on fraction of subreads (passes) detecting the call and fraction of subreads overlapping the call' step of the call-filtering pipeline. We next calculate germline variant sensitivity for the sample as the number of true-positive germline variant calls that pass the same filtering thresholds used in the 'Filters based on fraction of subreads (passes) detecting the call and fraction of subreads overlapping the call' step of the call-filtering pipeline, divided by the total number of true-positive germline variant calls. Each sample's dsDNA call counts are then corrected for sensitivity by dividing by that sample's calculated germline variant sensitivity. Each sample's ssDNA call counts are corrected by dividing by the square root of that sample's germline variant sensitivity, because the above dsDNA germline variant sensitivity estimate corrects for filters applied to both strands separately.

Finally, ssDNA and dsDNA burdens corrected for both trinucleotide context and sensitivity are calculated as the sum of the trinucleotide context- and sensitivity-corrected call counts divided by the number of interrogated bases (ssDNA burdens) or base pairs (dsDNA burdens). For all analyses and figures, unless otherwise specified, we use burden estimates corrected for both the full genome trinucleotide distribution and sensitivity.

The Poisson 95% confidence intervals of a sample's corrected burden were calculated as the corrected burden × [Poisson 95% confidence interval of raw call counts, calculated using the poisson.test function in R]/[raw call counts]. Weighted least-squares linear regressions of call burdens versus age were performed using the 'lm' function in R (via the ggplot[91] package), with weights equal to 1/[raw call counts].

HiDEF-seq estimate of fidelity for dsDNA mutations

The fidelity for dsDNA mutations was estimated for each sample as follows: (1) for each of the 192 possible trinucleotide contexts (that is, both central pyrimidine and central purine contexts), the number of single-strand calls at that context was divided by the total number of interrogated bases with that trinucleotide context to obtain a ssDNA call burden for that context; (2) for each central pyrimidine trinucleotide context, a dsDNA mutation error probability was calculated by multiplying the single-strand call burdens of the corresponding central pyrimidine and reverse-complement central purine trinucleotide contexts; and (3) all of the resulting central pyrimidine trinucleotide context dsDNA mutation error probabilities were summed. The main text reports the average fidelity across samples from healthy individuals, excluding sperm samples (as these have an outlier high ssDNA

C>T burden) and post-mortem samples processed with HiDEF-seq with A-tailing.

Comparison of HiDEF-seq and standard PacBio HiFi molecule characteristics

Standard PacBio HiFi raw subread data for comparison to HiDEF-seq (Fig. 1b and Extended Data Fig. 1d,f) were obtained from the Human Pangenome Reference Consortium (HPRC) public data repository[92] (samples HG02080, HG03098, HG02055, HG03492, HG02109, HG01442, HG02145, HG02004, HG01496, HG02083). Circular consensus sequences were derived from raw subreads using the same ccs version and ccs parameters used to analyse HiDEF-seq data.

Comparison of HiDEF-seq mutation burdens in sperm to paternally phased de novo mutation burdens

Paternally phased de novo mutation (DNM) burdens were calculated for each paternal age (in one-year intervals) from data published in a previous study of 2,976 trios[14] (supplementary files aau1043_datas5_revision1.tsv and aau1043_datas7.tsv), and using additional methodological details obtained from its associated study[93]. Paternally phased DNM burdens were first calculated for each child as [total number of paternally phased DNMs]/[fraction of the child's DNMs that were either paternally or maternally phased (which corrects for each child's phasing rate)] × [the Jónsson et al.[93] correction factor of 1.009 (which accounts for its false-positive and false-negative rate)]/[the Jónsson et al.[93] interrogated genome size of 2,682,890,000] (refs. 14,93). We then compare the dsDNA mutation burden of each HiDEF-seq sperm sample to the DNM burdens of children whose fathers' age at their birth is one year higher than the age at which the sperm sample was collected (to account for around 9 months difference between the father's age at conception and the child's birth).

Comparison of HiDEF-seq and NanoSeq call burdens and patterns

NanosSeq data were processed using the NanoSeq analysis pipeline v.3.2.1 (https://github.com/cancerit/NanoSeq) for chromosomes 1–22 and X (hs37d5 reference genome). NanoSeq dsDNA burdens corrected for trinucleotide context opportunities were obtained from the 'results.mut_burden.tsv' output file of the NanoSeq pipeline. NanoSeq ssDNA call burdens were calculated as the sum of the values in the 'results.mismatches.subst_asym.tsv' output file, divided by 2 × [the sum of the values in the 'results.mismatches.subst_asym.tsv' output file + the number of interrogated dsDNA base pairs obtained from the 'results.mut_burden.tsv' output file]. NanoSeq ssDNA call counts for each context were obtained from the 'results.SSC-mismatches-Pyrimidine.triprofiles.tsv' and 'results.SSC-mismatches-Purine.triprofiles.tsv' output files. Because the NanoSeq pipeline does not correct ssDNA calls for trinucleotide context opportunities, we compared the burdens of NanoSeq ssDNA calls in each context to the burdens of HiDEF-seq ssDNA calls that are also not corrected for trinucleotide context opportunities (that is, to HiDEF-seq burdens corrected only for sensitivity) (Fig. 1f,g and Extended Data Fig. 6b,c).

For more informative comparison of Poisson 95% confidence intervals of HiDEF-seq and NanoSeq (Fig. 1c,e,f and Extended Data Fig. 4a) despite a different number of interrogated bases (for ssDNA calls, or base pairs for dsDNA calls) measured by each method, for each sample, the number of calls of the method with the higher number of interrogated bases (or base pairs) was downsampled proportionally to the ratio of the number of interrogated bases of the two methods. The downsampled method's burden was then recalculated as the downsampled call count divided by the number of interrogated bases of the method with fewer interrogated bases, and the downsampled method's Poisson 95% confidence interval was recalculated using the downsampled number of raw call counts. This downsampling does not affect burden estimates, and it normalizes the confidence intervals of the two methods to reflect an equivalent number of interrogated bases (or base pairs). Confidence intervals before downsampling are provided in Supplementary Table 2.

Transcription level and transcription strand analysis of sperm HiDEF-seq ssDNA C>T calls

We obtained RNA-seq data of purified human spermatozoa from supplementary table 2 of ref. 94 ('Expression' sheet, average of the Control 1, Control 2, and Control 3 samples' fragments per kilobase of transcript per million mapped reads (FPKM) values) and annotated each gene that had non-zero expression with its expression quartile. We joined these data to the UCSC CHM13 v.1.0 genome browser 'CAT Gene + LiftOff Annotations V4' transcript annotation track using Ensembl gene IDs. We then annotated each ssDNA C>T call in HiDEF-seq sperm samples with the transcript expression data, and further annotated for each call if it was present on the transcribed or non-transcribed strand. We excluded from analysis the small number of calls overlapping transcripts expressed on both strands. We next calculated the sum of the lengths of transcripts in each expression quartile, excluding regions with transcripts expressed on both strands. We then normalized the number of ssDNA C>T calls in each quartile and each transcribed/non-transcribed strand category by the sum of the lengths of transcripts in that quartile. We then normalized these values for each transcribed and non-transcribed strand category by the sum of that category's values.

Signature analysis

Signature analysis for dsDNA mutations was performed using the 'sigfit' package[95] v.2.2, with input of raw mutation counts for each trinucleotide context, and the 'opportunities' parameter set to the ratio of the fractional abundance of each trinucleotide context in interrogated bases of that sample versus the fractional abundance of that trinucleotide context in the human reference genome. The correction for trinucleotide context opportunities performed above for burden analyses used the fractional abundance of trinucleotides in CHM13 v.1.0, but the correction for trinucleotide context opportunities performed here for signature analysis and figures used the fractional abundance of trinucleotides in the full GRCh37 genome (for both nuclear and mitochondrial genome analyses and figures) so that the obtained spectra and signatures can be compared to standard COSMIC signatures. The 'plot_gof' function was used to determine the optimal number of signatures to extract. As COSMIC SBS1 was not well separated from other signatures during de novo extraction[96], we used the 'fit_extract_signatures' function to fit SBS1 while simultaneously extracting additional signatures de novo. De novo extracted signatures were compared to the COSMIC SBS v.3.2 catalogue[23] to identify the most similar known signature by cosine similarity. To obtain more accurate estimates of signature exposures, the fitted COSMIC SBS signature and the extracted signatures were then re-fit back to the mutation counts using the 'fit_signatures' function, along with correction for trinucleotide context opportunities. SBS5 is a ubiquitous clock-like signature[23], and often de novo extraction produced more than one signature with weak or moderate similarity to SBS5, for example, both SBS5 and SBS40 (cosine similarity = 0.83) or both SBS3 and SBS40 (cosine similarity = 0.88). In these cases, we either reduced the number of de novo extracted signatures so that only one of these similar signatures was extracted, or we instructed 'fit_extract_signatures' to fit both COSMIC SBS1 and COSMIC SBS5.

ssDNA signatures were extracted by taking advantage of sigfit's ability to analyse 192-trinucleotide context mutational spectra that distinguish transcribed versus untranscribed strands. Instead, we use this feature to distinguish central pyrimidine versus central purine contexts. We do this by arbitrarily setting central pyrimidine and central purine ssDNA calls to the transcribed and untranscribed strands, respectively (by setting the strand column to '−1' for all calls that are input into sigfit's 'build_catalogues' function, without collapsing central

pyrimidine and central purine contexts). We then extract ssDNA signatures as described above for dsDNA signatures, with correction for trinucleotide context opportunities. Cosine similarities of ssDNA and dsDNA signatures are calculated after projecting ssDNA signatures to 96-central pyrimidine contexts, which is performed by summing values of central pyrimidine contexts with values of their reverse-complement central purine contexts.

To help to qualify the significance of cosine similarities, we performed simulations of random 96-element and 192-element number vectors (n = 10,000 random vector pairs each), which showed that 5.9%, 0.06% and 0% of random 96-context cosine similarities are above cut-offs of 0.8, 0.85 and 0.9, respectively, and 1.2%, 0% and 0% of random 192-context cosine similarities are above cut-offs of 0.8, 0.85 and 0.9, respectively. Thus, for 96-context comparisons (that is, dsDNA and projected ssDNA to dsDNA comparisons), we use the qualitative terms 'weak similarity' for 0.8 ≤ cosine similarity < 0.85, 'moderate similarity' for 0.85 ≤ cosine similarity < 0.9, and 'strong similarity' for cosine similarity ≥ 0.9. For 192-context comparisons (that is, ssDNA to ssDNA comparisons), we use the terms 'moderate similarity' for 0.8 ≤ cosine similarity < 0.85 and 'strong similarity' for cosine similarity ≥ 0.85.

Replication strand asymmetry (fork polarity) analysis

ENCODE replication timing (Repli-seq) data[97] (wavelet-smoothed signal) were obtained from the UCSC Genome Browser[80] (hg19) for the lymphoblastoid cell lines GM12878, GM06990, GM12801, GM12812 and GM12813. We calculated the average of the Repli-seq signal (higher values indicate earlier replication) across these samples at each position, and then lifted over the data to CHM13 v.1.0. For each analysed HiDEF-seq call, we calculated the fork polarity[98] as the slope versus position of the Repli-seq data points spanning −5 to +5 kb from the call using the 'lm' function in R. Positive and negative fork polarities indicate the genome non-reference (−) strand is synthesized more frequently in the leading- and lagging-strand direction, respectively. This was also performed for a set of 50 iterations of 1,000 randomly selected genomic positions with either the sequence or the reverse complement of the sequence corresponding to the trinucleotide context being analysed (that is, AGA or TCT for *POLE* samples). We next calculated the fork polarity quantile values at quantiles ranging from 0 to 1.0 in 0.1 increments, and then for each of these quantile bins (combining 0.4–0.5 and 0.5–0.6 quantile bins into one bin, as these span fork polarity 0), we counted the number of loci whose sequence is AGA in the genome non-reference (−) strand and the number of loci whose sequence is AGA in the reference genome (+) strand. Loci without annotated Repli-seq data were excluded. Next, for each genome strand, we calculated normalized call counts by dividing the quantile bin call counts by the total number of calls in that strand. For each of the nine quantile bins, we then calculated the 'strand ratio' as the ratio of non-reference to reference strand normalized call counts. We also calculated this strand ratio for positive and negative fork polarities (that is, two bins rather than nine quantile bins), as there were not enough ssDNA calls in individual quantile bins for analysis. Analyses were also repeated after excluding loci within genic regions annotated in the CHM13 v.1.0 LiftOff Genes V2 annotation obtained from the UCSC Genome Browser.

Kinetics analysis

Signatures of sequencing polymerase kinetics have been previously identified for diverse base modifications in synthetic oligonucleotides, and they have been used to detect a small number of base modifications in genomic DNA such as cytosine methylation[43,99]. However, this approach has not yet been used to detect uracil-species in genomic DNA with single-molecule fidelity. We performed the kinetics analysis as follows.

For each sample, consensus sequences for each strand were created using pbccs v.6.4.0 (Pacific Biosciences) with the parameters: --by-strand --hifi-kinetics --min-rq 0.99 --top-passes 0. pbccs v.6.4.0 was used because, with these parameters, it outputs consensus kinetics values for each strand separately, which previous versions of pbccs do not. Consensus sequence reads were then aligned to the CHM13 v.1.0 reference genome with pbmm2 with the parameters '--preset CCS --sort'.

Next, we extracted the list of ssDNA C>T sequence calls in the 72 °C heat-treated blood DNA and the sperm samples (profiled by HiDEF-seq with nick ligation). Owing to the very high number of ssDNA C>T calls in blood DNA samples that were heat treated in water-only or Tris-only buffer, for these samples, we selected a random subset of 800 calls. We then extracted from these samples and from 85 other HiDEF-seq samples all of the consensus reads that overlapped the C>T call positions, from the strand synthesized by the sequencing polymerase opposite to the strand on which the call is present in the molecule. As kinetics is affected by sequence context[43], this enables calculation of differences in kinetics between molecules with and without the event within the same sequence context. We next performed kinetic analyses of IPD and PW. Kinetics values (IPD or PW, reported by the sequencing instrument at a 10 ms frame rate) for each consensus read were transformed into units of time (seconds) and normalized by the average kinetics values of all bases in the consensus read to correct for baseline sequencing kinetics differences between molecules. For each C>T call, we extracted the kinetics values of all overlapping reads for ±30 bp flanking the event position relative to the reference genome coordinates using each read's CIGAR value to account for insertions or deletions in the read relative to the reference genome. Next, for each C>T call, we calculated the ratio of kinetics values for each base position by dividing the kinetics values (IPD or PW) of the molecule with the call by the weighted average kinetics values of molecules without the call (the weighted average weights by each molecule's number of passes; that is, its 'ec' tag value). Finally, for each flanking and mutant base position, we calculated the average and s.e.m. of the kinetics value ratios across all C>T calls of each sample or sample set of interest. The same kinetic analysis was performed for dsDNA C>T mutation calls (that is, bona fide cytosine to thymine double-strand mutations) in non-heat-treated blood DNA, 56 °C and 72 °C heat-treated blood DNA, sperm, kidney, and liver samples (all profiled by HiDEF-seq with nick ligation), for the strands synthesized by the sequencing polymerase opposite the strand containing the C>T mutation; this shows the kinetic profile of true C>T changes, as a comparator for C>T calls arising from cytosine damage. Note that the dsDNA C>T mutations used for this kinetics analysis were called with the same thresholds used for ssDNA C>T calls. Both these ssDNA and dsDNA analyses were additionally conducted after randomization of labels among molecules with and without the C>T call to confirm that the kinetic signal was specific to molecules with the C>T call. The kinetic profile heat map and clustering were performed using the ComplexHeatmap R package[100].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Sequencing data generated in this study (FASTQ files for Illumina sequencing; subreads BAM files for PacBio data) are available at the NCBI database of Genotypes and Phenotypes under accession code phs003604 (all of the samples except those from the International Replication Repair Deficiency Consortium and participants D1 and D2) and at the European Genome–Phenome Archive under accession number EGAS50000000318 (samples from the International Replication Repair Deficiency Consortium). Sequencing data of participants D1 and D2 were not deposited in these databases due to consent limitations. Accession IDs of specific samples are provided in Supplementary Table 1.

Code availability

The source code for the HiDEF-seq analysis pipeline is available at GitHub (https://github.com/evronylab/HiDEF-seq), and the version used for this manuscript (v.1.1) is archived in Zenodo (https://doi.org/10.5281/zenodo.10898439).

54. Agarwal, A., Gupta, S. & Sharma, R. in *Andrological Evaluation of Male Infertility: A Laboratory Guide* (eds Agarwal, A. et al.) 101–107 (Springer, 2016).
55. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
56. Wu, H., de Gannes, M. K., Luchetti, G. & Pilsner, J. R. Rapid method for the isolation of mammalian sperm DNA. *BioTechniques* **58**, 293–300 (2015).
57. Jenkins, T. G., Liu, L., Aston, K. I. & Carrell, D. T. Pre-screening method for somatic cell contamination in human sperm epigenetic studies. *Syst. Biol. Reprod. Med.* **64**, 146–155 (2018).
58. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
59. Heng, L. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
60. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
61. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
62. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
64. R Core Team. *R: A Language and Environment for Statistical Computing* (2021).
65. Martin, M., Hervé, P., Valerie, O. & Nathaniel, H. Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix (2020).
66. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
67. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).
68. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
69. Jianfeng, L. configr: an implementation of parsing and writing configuration file (2020).
70. Ching, T. qs: quick serialization of R objects https://CRAN.R-project.org/package=qs (2021).
71. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
72. Milton, S. & Wickham, H. magrittr: a forward-pipe operator for R (2020).
73. Wickham, H., Hester, J. & Bryan, J. readr: read rectangular text data (2022).
74. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: a grammar of data manipulation (2021).
75. Lee, S., Cook, D. & Lawrence, M. plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**, 4 (2019).
76. Wickham, H. stringr: simple, consistent wrappers for common string operations (2019).
77. Eddelbuettel, D. digest: create compact hash digests of R objects (2021).
78. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
79. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
80. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
81. Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P. & Flicek, P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* **30**, 1008–1009 (2014).
82. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
83. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
84. Hunt, M. et al. REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14**, R47 (2013).
85. Ponstingl, H. & Ning, Z. SMALT - a new mapper for DNA sequencing reads [poster]. *F1000Posters* **1**, 313 (2010).
86. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
87. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
88. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2015).
89. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).
90. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
91. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
92. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
93. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519 (2017).
94. Zhu, C.-H. et al. Investigation of the mechanisms leading to human sperm DNA damage based on transcriptome analysis by RNA-seq techniques. *Reprod. BioMed. Online* **46**, 11–19 (2023).
95. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at *bioRxiv* https://doi.org/10.1101/372896 (2020).
96. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
97. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107**, 139–144 (2010).
98. Seplyarskiy, V. B. et al. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26**, 174–182 (2016).
99. Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
100. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
101. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
102. Freudenthal, B. D., Beard, W. A., Shock, D. D. & Wilson, S. H. Observing a DNA polymerase choose right from wrong. *Cell* **154**, 157–168 (2013).
103. Verderio, P. et al. External quality assurance programs for processing methods provide evidence on impact of preanalytical variables. *New Biotechnol.* **72**, 29–37 (2022).

Acknowledgements This work was supported by grants from the NIH Common Fund (UG3NS132024 – Somatic Mosaicism across Human Tissues Network, G.D.E.; DP5OD028158, G.D.E.), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R21HD105910; G.D.E. and J.E.S.), the Sontag Foundation (G.D.E.), the Pew Charitable Trusts (G.D.E.) and the Jacob Goldfield Foundation (G.D.E.). Sequencing performed at the New York University (NYU) Grossman School of Medicine Genome Technology Center was supported in part by the National Cancer Institute (P30CA016087) and a National Institutes of Health Shared Instrumentation Grant (1S10OD023423-01). The computational work was supported in part by the New York University Information Technology High Performance Computing resources, services and staff expertise, and by the New York University Grossman School of Medicine High Performance Computing Core. U.T. was supported by a Stand Up To Cancer–Bristol-Myers Squibb Catalyst Research Grant (SU2C-AACR-CT-07-17), SickKids Foundation donors Harry and Agnieszka Hall, Meagan's Walk (MW-2014-10), BRAINchild Canada, the LivWise Foundation, the Canadian Institutes for Health Research (CIHR; grant 108188) and a Canadian Cancer Society/CIHR/Brain Canada Spark Grant (Spark-21, 707089). J.E.S. was supported by the Damon Runyon Cancer Research Foundation, the Vinney Family Scholars Award and the Bristol Myers Squibb Foundation. M.G.-P. was supported by NIH grants T32AG052909 and F32AG076287. We thank B. Neel, H. Klein and A. Chakravarti (NYU Grossman School of Medicine) for discussions; D. Dimartino and P. Zappile (Genome Technology Center at NYU Grossman School of Medicine) for assistance with sequencing; M. Fridrikh, N. Francoeur and R. Sebra (Genomics Core Facility at the Icahn School of Medicine at Mount Sinai) for assistance with sequencing; S. Wang (NYU Information Technology) for assistance with high-performance computing; and the NIH NeuroBioBank and its staff at the University of Maryland (R. Johnson) for providing human tissues.

Author contributions G.D.E. conceived the project. G.D.E., M.H.L., B.M.C., U.C. and J.E.S. designed the experiments. U.T., V.B., L.S., N.M.N., T.P. and R.E.B. collected some of the samples. E.L., D.R. and A.-B.S. recruited research participants for sperm samples. D.R. prepared ZyMot sperm samples. M.H.L., R.C.B., A.S., Z.R.M., C.A.L., T.K.T. and G.D.E. prepared tissues and cell samples. M.H.L., B.M.C. and U.C. performed technology development experiments. M.H.L. and B.M.C. prepared HiDEF-seq sequencing libraries. M.G.-P. prepared NanoSeq libraries. M.H.L., B.M.C. and E.C.B. performed other experiments. J.R.W. assisted with interpretation of cytosine deamination data. G.D.E. created the computational pipeline with input from U.C. M.H.L., B.M.C. and G.D.E. performed the analysis. M.H.L. and G.D.E. wrote the initial manuscript, with input from B.M.C. and J.E.S. All of the authors contributed to the final manuscript.

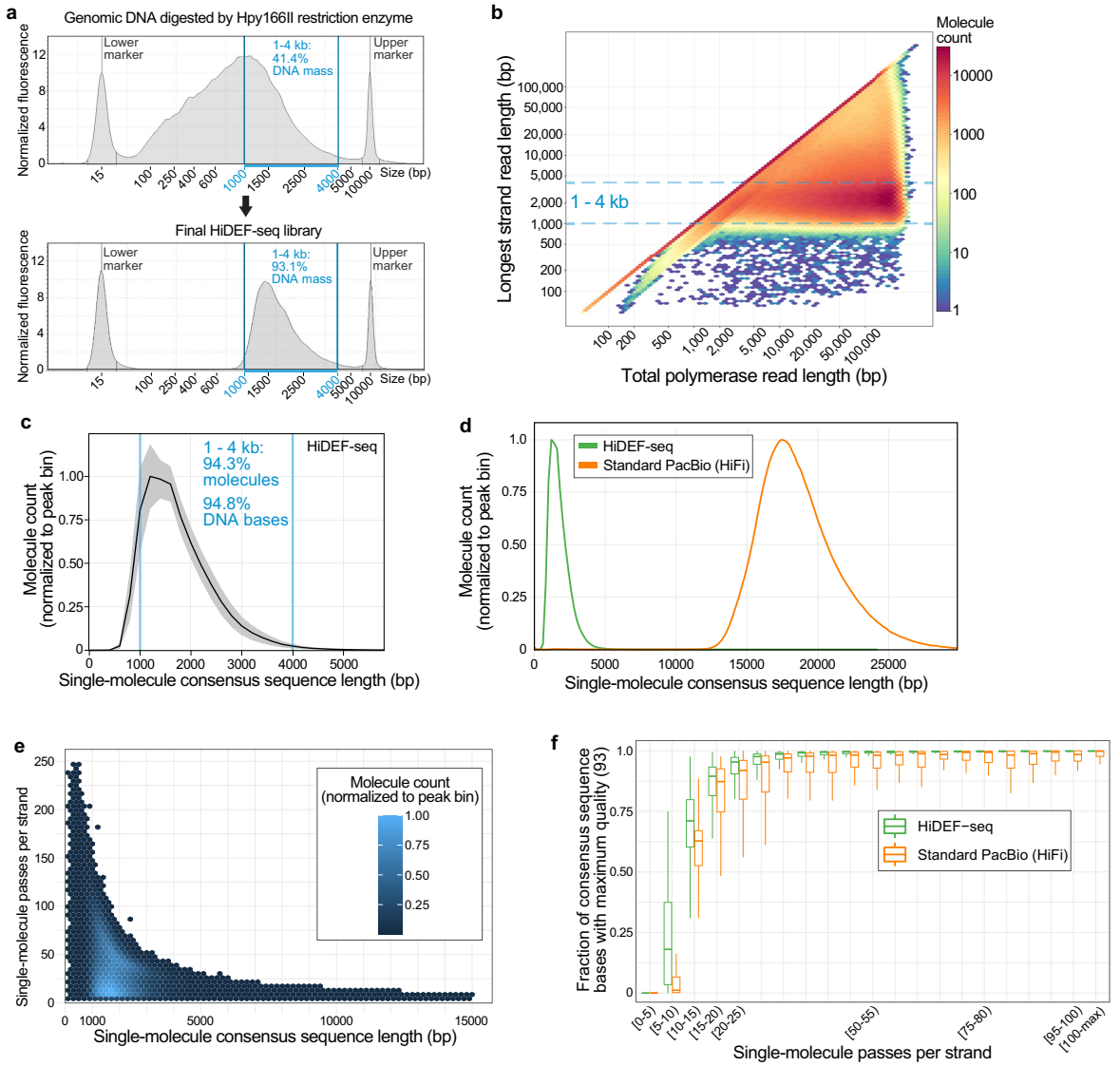
Competing interests A patent application for HiDEF-seq has been filed (G.D.E.). G.D.E. owns equity in DNA sequencing companies (Illumina, Oxford Nanopore Technologies, and Pacific Biosciences). The other authors declare no competing interests.

Additional information

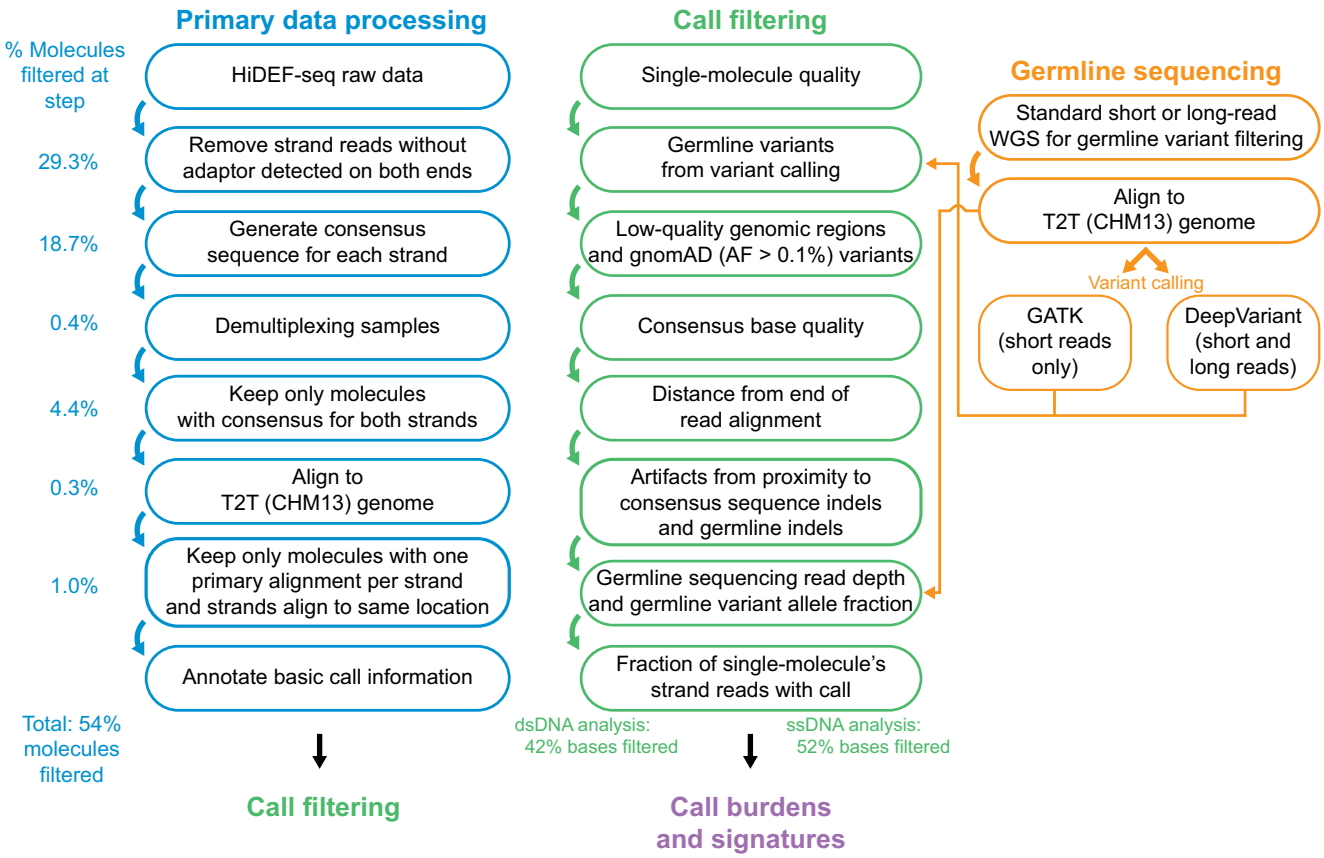
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07532-8.

Correspondence and requests for materials should be addressed to Gilad D. Evrony.

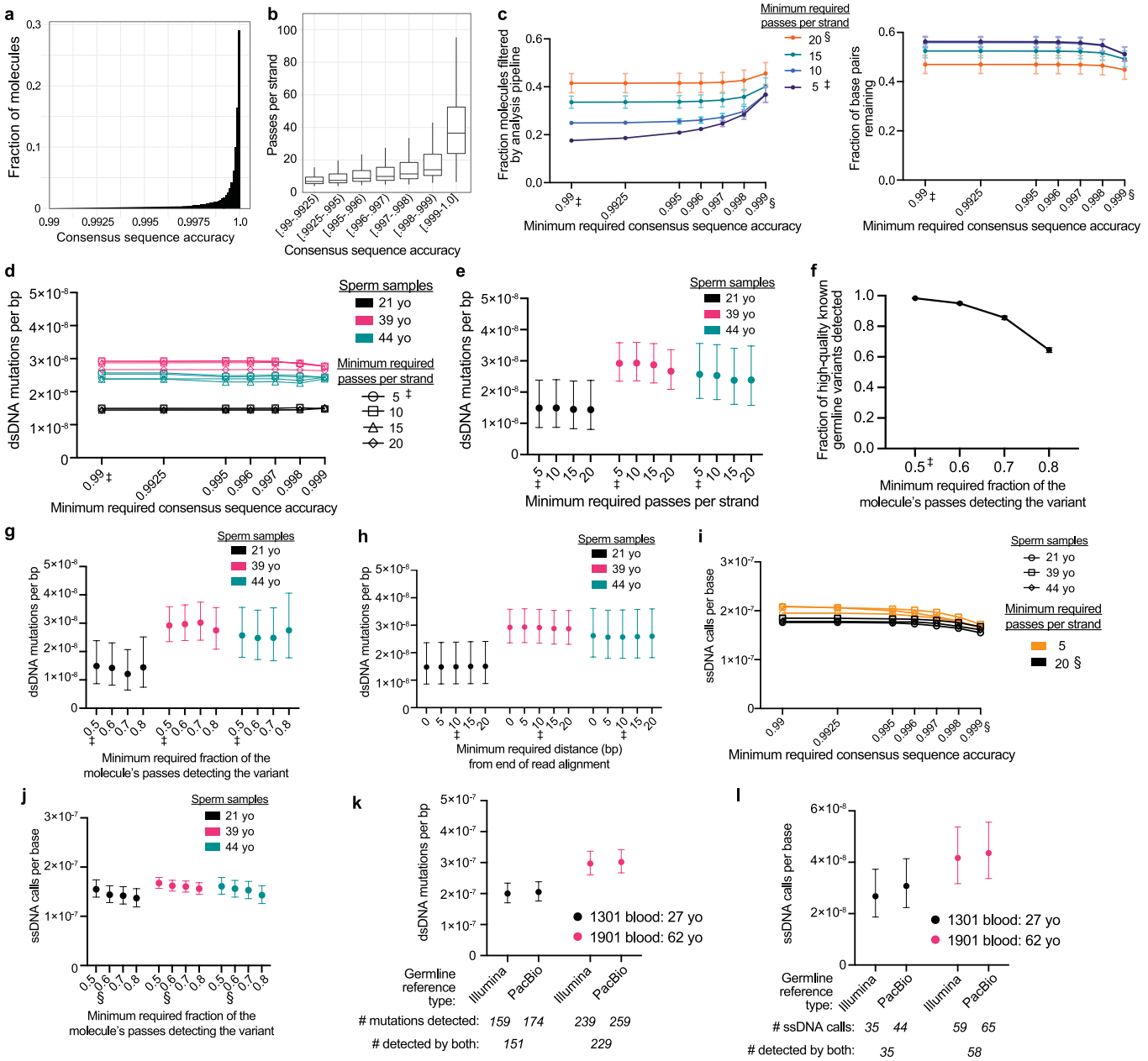
Peer review information *Nature* thanks Francesca Storici and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | HiDEF-seq library preparation and sequencing metrics. a, Representative DNA sizing electropherogram after Hpy166II restriction enzyme digestion (top) and after completion of the HiDEF-seq library preparation, which removes fragments <1 kb (bottom). **b**, Two-dimensional histogram of all molecules from a representative HiDEF-seq sequencing run of each molecule's longest strand read length (bp, base pairs) versus its total polymerase read length (PRL). Dashed line signifies the expected strand length distribution. The red diagonal line reflects 18% of molecules with <1 strand pass, which is typical in PacBio sequencing. **c**, Histogram (200 bp bins) for representative HiDEF-seq samples (n = 51) of molecule consensus sequence lengths (i.e., molecule sizes). Line and shaded region show average and standard deviation, respectively, across samples for each bin. The average of these samples' median lengths is 1.7 kilobases (kb). **d**, Histogram as in panel (c), showing HiDEF-seq (n = 51 representative samples) yields smaller molecule lengths than standard PacBio (HiFi) samples (n = 10 samples). The average of samples' median lengths are 1.7 kb and 18.3 kb for HiDEF-seq and HiFi, respectively. **e**, Two-dimensional histogram of the number of passes (bin width of 5 passes) vs. consensus sequence lengths (bin width of 200 bp) for molecules from the 51 representative HiDEF-seq samples plotted in panels (c,d). Bins are coloured if there is at least one molecule in the bin. **f**, Box plots of the fraction of a molecule's consensus sequence bases (average of forward and reverse strands) that have the maximum predicted quality (quality=93, as predicted by ccs, Methods) versus the number of passes per strand, across all molecules of the same samples included in panels (c-e). Note: 93 is the quality required for HiDEF-seq analysis. This plot illustrates that the number of passes is a key determinant of consensus quality in both HiDEF-seq and HiFi. **b**, Plot generated by SMRT Link (Pacific Biosciences) software. **c**-**e**, The single-molecule consensus sequence length is the average of the forward and reverse strand lengths. Bin values are normalized to the bin with the highest molecule count. **e**,**f**, The number of passes per strand is the average of the forward and reverse strand 'ec' tags (Methods). **c**-**f**, Plots show data of HiDEF-seq molecules that are output by the primary data processing step of the HiDEF-seq analysis pipeline and standard PacBio HiFi molecules that are output by the ccs HiFi pipeline (Methods). **f**, Box plot: middle line, median; boxes, 1st and 3rd quartiles; whiskers, the maximum/minimum values within 1.5 x interquartile range. X-axis: square brackets and parentheses signify inclusion and exclusion of interval endpoints, respectively.

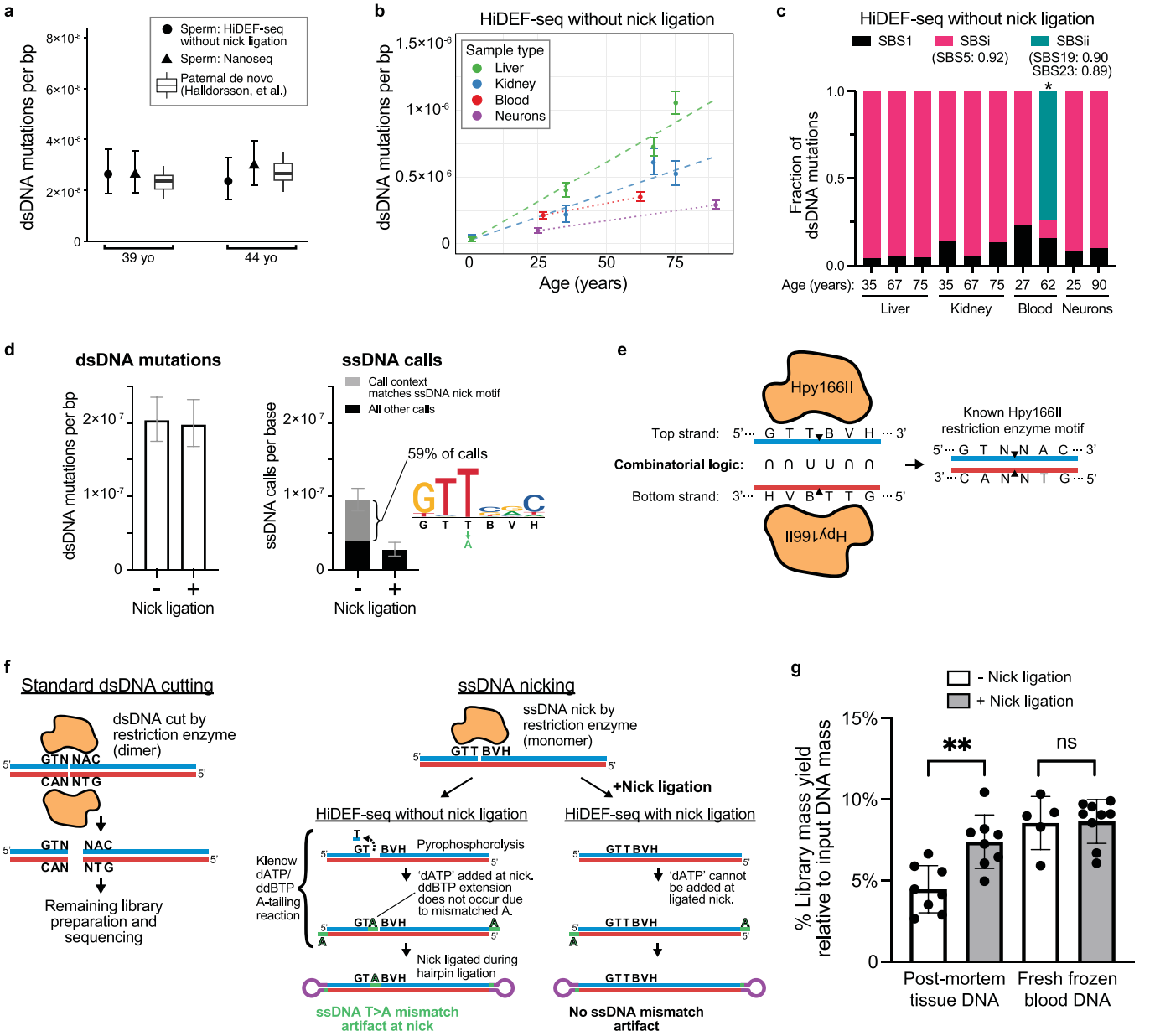


Extended Data Fig. 2 | Schematic of analysis pipeline. Primary data processing (blue) is followed by call filtering (green) along with germline sequencing analysis (orange), which is then followed by call burden and signature analysis (purple). See Methods for full details. On the left of primary data processing steps are the average percentage of molecules filtered by each step across 17 representative HiDEF-seq sequencing runs. Approximately half of molecules filtered by the 'Generate consensus sequence' step are molecules with less than 3 full-length passes (default setting of the ccs tool that creates consensus sequences), and the other half are due to molecules with read quality ('rq' tag) <0.99. At the end of the call filtering steps are listed the percentage of bases filtered by all the call filtering steps, calculated out of the total bases of molecules that pass primary data processing, for the same 17 representative HiDEF-seq sequencing runs. The filter for 'low-quality genomic regions and gnomAD variants with allele frequency (AF) > 0.1% in the population' covers approximately 15% and 7% of the genome when using Illumina and PacBio germline sequencing data, respectively (i.e., when PacBio germline sequencing data is used, the pipeline uses less restrictive filters due to fewer genome alignment errors and artifacts). WGS, whole-genome sequencing.



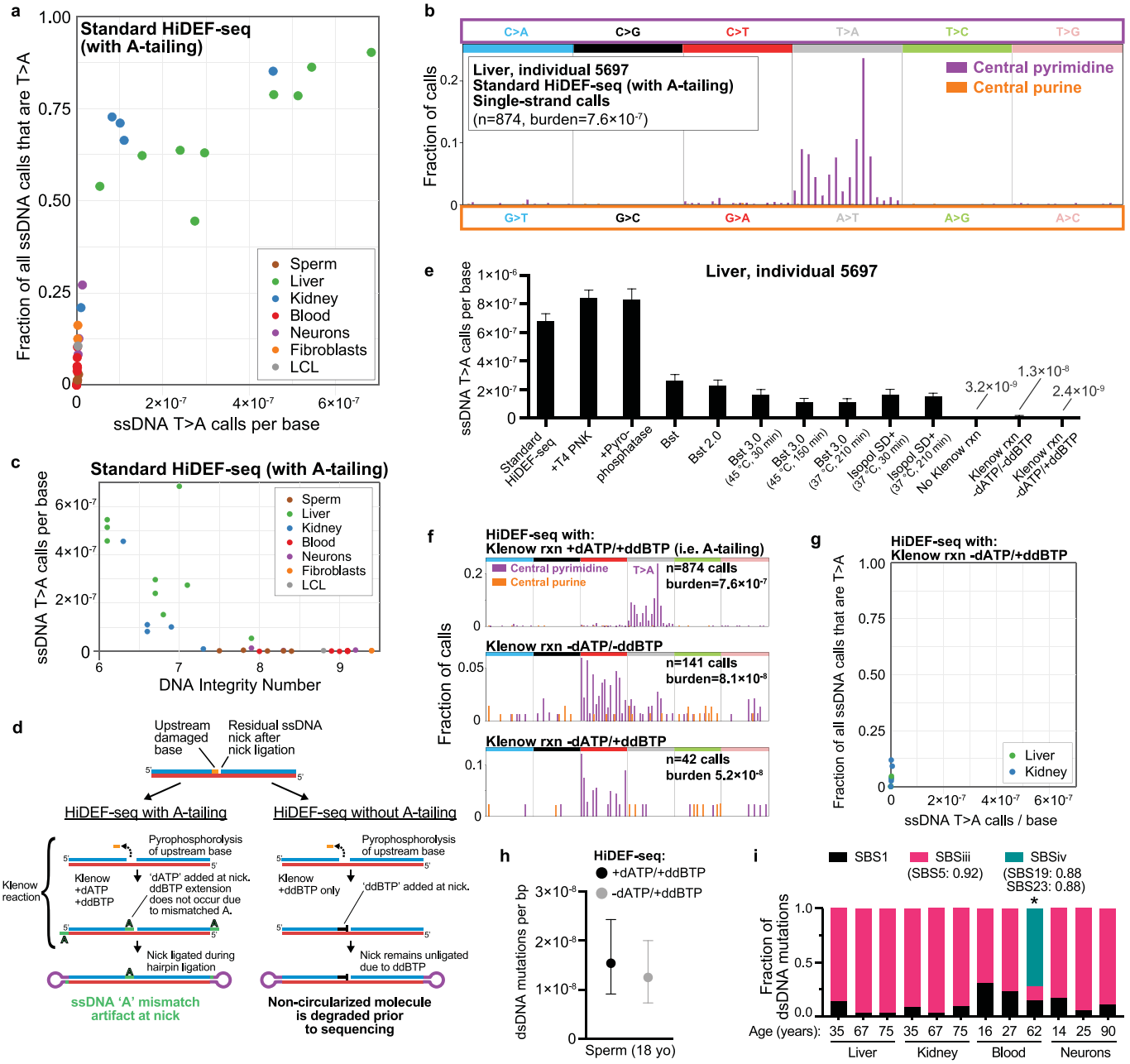
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Analysis thresholds and comparison of analyses using short- versus long-read germline sequencing. a, Histogram of predicted consensus sequence accuracy ('rq' tag, bin width=0.0001) for DNA molecules that pass primary data processing steps from 3 representative sperm samples profiled by HiDEF-seq (with nick ligation) (21yo: SPM-1002; 39yo: SPM-1004; 44yo: SPM-1020; yo, years old). Note, these are consensus sequence accuracies predicted by the ccs consensus calling software (Methods), which are used to filter low-quality molecules, but these accuracies do not reflect the true accuracy that is significantly higher. **b**, Box plot of passes per strand for different consensus sequence accuracy bins, for molecules from the 3 samples included in the prior panel, showing that higher minimum accuracies select for molecules with higher numbers of passes. **c**, Fraction of post-primary data processing molecules that are filtered (left plot) and fraction of post-primary data processing base pairs that remain for interrogation (right plot) using different minimum passes per strand and consensus sequence accuracy thresholds. Values show average of the 3 samples included in the prior panels, after completing all steps of the mutation filtering pipeline. **d**,**e**, dsDNA mutation burdens for the 3 samples included in the prior panels using different minimum passes per strand and consensus sequence accuracy thresholds. Panel (e) shows data from (d) at consensus accuracy of 0.99 with Poisson 95% confidence intervals. These data illustrate stability of dsDNA mutation burden estimates at broad thresholds using sperm as the most stringent test of fidelity. **f**, Fraction of high-quality, known heterozygous germline variants detected using different minimum required fraction of molecule passes (i.e., subreads) that detect the variant (filter applied separately to each strand). This value is used for sensitivity correction (Methods). Values show average of the 3 samples included in prior panels. **g**,**h**, dsDNA mutation burdens for the 3 samples included in the prior panels using different minimum required fraction of molecule passes that detect the variant (filter applied separately to each strand), after correcting for sensitivity (g), and using different minimum required distances from the end of the read (h). Panel (g) illustrates that correcting for sensitivity maintains stable burden estimates. The analysis pipeline requires a minimum of 10 bp from the ends of reads to remove rare alignment artifacts, although this does not significantly alter burden estimates. **i**, ssDNA call burdens for the 3 sperm samples included in the prior panels using different minimum passes per strand and consensus sequence accuracy thresholds. Plot shows a small decrease in ssDNA call burdens with a higher minimum required passes per strand at low consensus sequence accuracy thresholds, and convergence to similar burdens at high consensus sequence accuracy thresholds. Data shown with minimum fraction of 0.5 molecule passes that detect the variant. **j**, ssDNA call burdens for the 3 sperm samples included in the prior panels using different minimum required fraction of molecule passes that detect the variant, after correcting for sensitivity. Data shown with a minimum consensus sequence accuracy of 0.999 and a minimum of 20 passes per strand. **k**,**l**, Concordant dsDNA mutation and ssDNA call burdens obtained by HiDEF-seq using short-read (Illumina) or long-read (PacBio, Pacific Biosciences) germline sequencing during analysis, for two samples (1301 and 1901 blood). **a-d**,**i**, Consensus sequence accuracies are the average of forward and reverse strand accuracies. **b**, Box plot: middle line, median; boxes, 1st and 3rd quartiles; whiskers, the maximum/minimum values within 1.5 x interquartile range. X-axis: square brackets and parentheses signify inclusion and exclusion of interval endpoints, respectively. **c-e**,**i**, Threshold for minimum required passes per strand is applied to both strands. **c-j**, The symbols ‡ and § mark the final thresholds chosen for dsDNA and ssDNA analyses, respectively. **c**,**f**, Error bars: standard deviation; note, panel (f) error bars are small and therefore not well visualized. **d,e,g-l**, Mutation and call burdens are corrected for sensitivity and trinucleotide context opportunities of the full genome relative to interrogated bases (Methods). **e**,**g,h**,**j-l**, Dots and error bars: point estimates and their Poisson 95% confidence intervals.



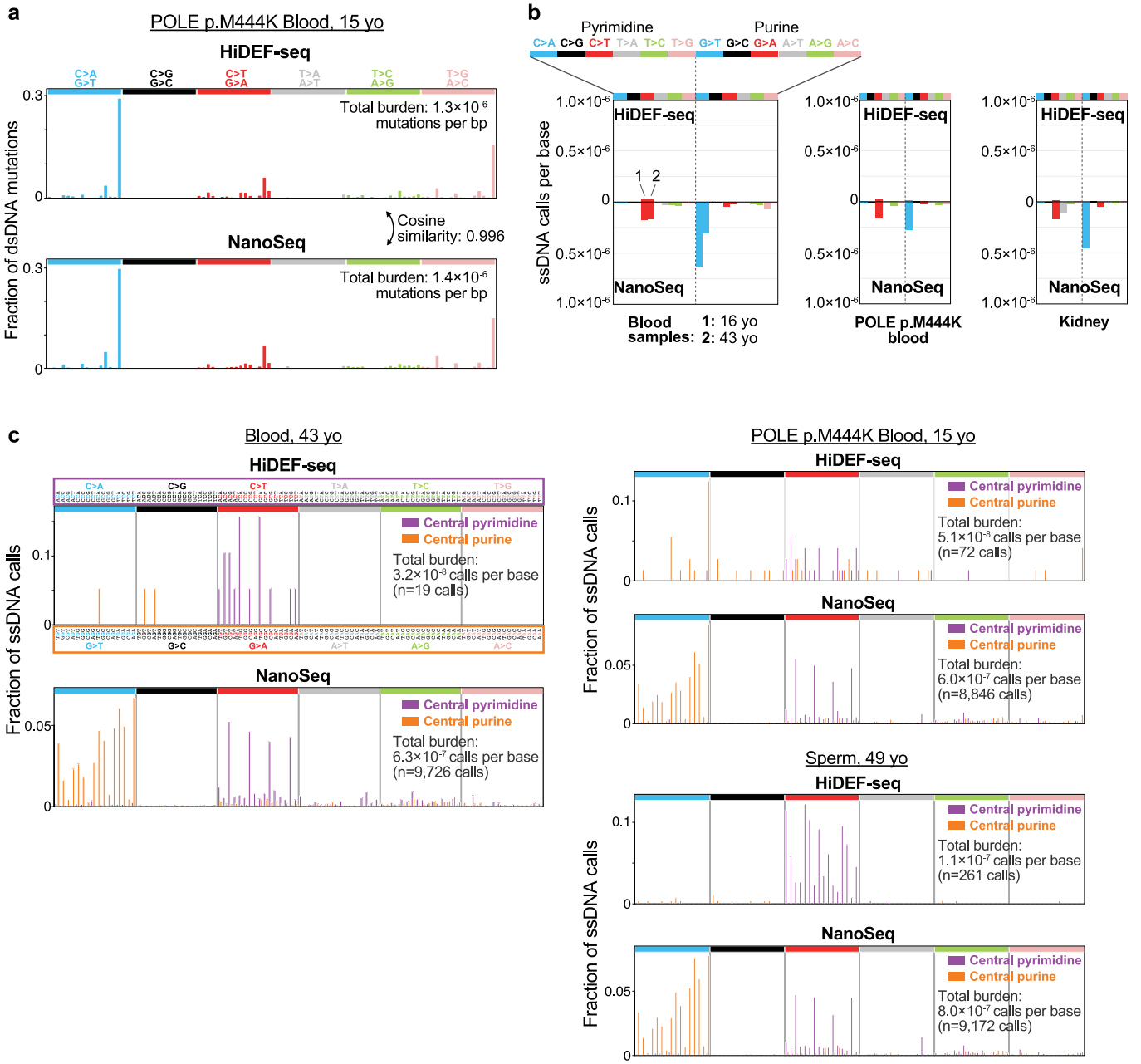
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | dsDNA mutation burdens of HiDEF-seq without ssDNA nick ligation and removal of ssDNA artifacts by ssDNA nick ligation. **a**, dsDNA mutation burdens in two sperm samples (left to right: SPM-1004, SPM-1020) profiled by HiDEF-seq without ssDNA nick ligation and by NanoSeq, compared for each age (yo, years old) to paternally phased de novo mutations in children from a prior study of 2,976 trios[14]. See Fig. 1c for sperm samples profiled by HiDEF-seq with nick ligation. **b**, dsDNA mutation burdens versus age, measured by HiDEF-seq without nick ligation (see Fig. 1d for samples profiled by HiDEF-seq with nick ligation). Dashed lines (liver, kidney): weighted least-squares linear regression. Dotted lines (blood, neurons): these only connect two data points to aid visualization of burden difference, since regression cannot be performed with two samples. **c**, Mutational signature contribution to dsDNA mutations detected in samples profiled by HiDEF-seq without nick ligation (see Extended Data Fig. 5i for samples profiled by HiDEF-seq with nick ligation). All samples, except blood from a 62-year-old individual with a history of kidney disease (1901, asterisk), were jointly analysed with fitting of SBS1 and de novo extraction of one additional signature SBSi (Methods). The blood sample of the 62-year-old was analysed separately together with 5 other HiDEF-seq (with nick ligation) blood samples from this individual, due to identification of an additional signature SBSii with strong and moderate similarity to SBS19 and SBS23, respectively. Analysis of samples grouped by tissue type, excluding the 62-year-old blood sample, produced similar results. For de novo extracted signatures (SBSi and SBSii), the cosine similarities to the most similar COSMIC signatures are shown in parentheses. Sperm samples and kidney and liver samples from an infant (1443) were not included here since the number of mutations is too low for reliable signature extraction. **d**, Burdens of dsDNA mutations (left plot) and ssDNA calls (right plot) of a blood sample (individual 1301) measured by HiDEF-seq without versus with nick ligation. Nick ligation eliminates T > A ssDNA artifacts that match the illustrated GTTBVH motif. The motif was derived using the ggseqlogo R package (ref. 101) using all ssDNA T > A calls from the sample profiled by HiDEF-seq without nick ligation. Grey bar is calls matching the motif with log-odds score > 2 calculated with the score_match function of the universalmotif R package. **e**,**f**, Proposed mechanism for the GTTBVH motif of ssDNA artifactual calls in HiDEF-seq without ssDNA nick ligation. The known GTNNAC motif of the Hpy166II restriction enzyme used in HiDEF-seq may arise if Hpy166II operates as a dimer (cut sites signified by triangles) with each monomer binding opposite strands, and the GTTBVH motif is due to intersection (∩) and union (U) combinatorial logic for the outer and inner 2 bases, respectively (e). Without nick ligation, ssDNA GT[T > A]BVH artifactual calls may arise from rare Hpy166II monomer nicking events, pyrophosphorolysis of the 'T' upstream of the nick, and addition of a mismatched 'A' during the Klenow dATP/ddBTP A-tailing reaction. Further extension with ddBTP does not occur due to the mismatch (ref. 102). This process is prevented in HiDEF-seq by nick ligation. **g**, Nick ligation increases HiDEF-seq library yield by 66% for post-mortem tissues, likely by repairing nicks in the original input DNA so that the molecules are not eliminated in the final nuclease treatment step. Bars show average yield for each group; number of samples per group (left to right): 8, 8, 5, 9 (**, p = 0.002; ns, not significant; two-sided unpaired t-test). **a**, Box plots: middle line, median; boxes, 1st and 3rd quartiles; whiskers, 5% and 95% quantiles. For each sample, HiDEF-seq and NanoSeq confidence intervals were normalized to reflect an equivalent number of interrogated base pairs (Methods). **a**,**b**,**d**, Error bars: Poisson 95% confidence intervals. **g**, Error bars: standard deviation.

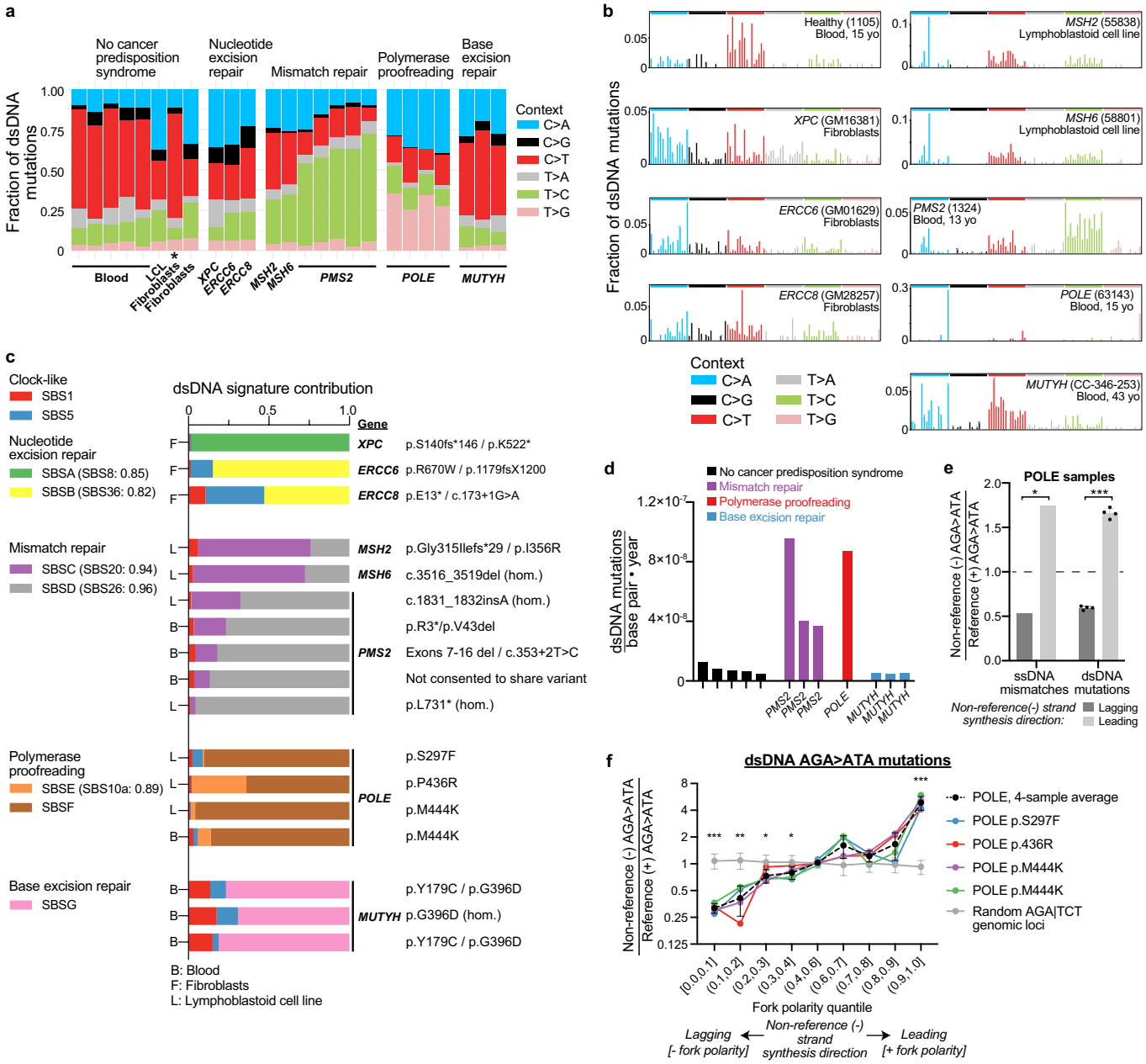


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | HiDEF-seq without A-tailing removes ssDNA artifacts of post-mortem tissues with fragmented DNA. a, Fraction of ssDNA calls that are T > A (corrected for trinucleotide context opportunities) versus the ssDNA T > A burden in all samples profiled by HiDEF-seq with A-tailing (i.e., Klenow reaction +dATP/+ddBTP) from healthy individuals and cell lines (i.e., excluding cancer predisposition syndromes). Post-mortem kidney and liver consistently have the highest fraction of ssDNA calls that are T > A. **b**, ssDNA call spectrum for a liver sample profiled by HiDEF-seq with A-tailing exhibiting a high ssDNA T > A burden ($6.8{\cdot}10^{-7}$ T > A burden; $7.6{\cdot}10^{-7}$ total ssDNA call burden), corrected for trinucleotide context opportunities. Parentheses show total number of calls. **c**, Correlation between ssDNA T > A artifact burden and the input DNA's DNA Integrity Number measured by TapeStation electrophoresis (ref. 103) across all samples profiled by HiDEF-seq with A-tailing from healthy individuals and cell lines (i.e., excluding cancer predisposition syndromes). Lower DNA Integrity Number corresponds to more fragmented DNA. **d**, Proposed mechanism for the ssDNA T > A artifact calls in fragmented DNA when performing HiDEF-seq with A-tailing and its prevention in HiDEF-seq without A-tailing. **e**, Modifications of the HiDEF-seq protocol to eliminate ssDNA T > A artifacts in fragmented DNA. All trials were from the same DNA extraction aliquot (liver from individual 5697). See Methods for details. PNK, polynucleotide kinase; Bst, Bst large fragment; min, minutes. **f**, ssDNA call spectra for three of the samples shown in panel (e): standard HiDEF-seq with A-tailing (top, same spectrum as panel (b)), HiDEF-seq with a Klenow reaction that does not contain dATP nor ddBTP (middle), and HiDEF-seq with a Klenow reaction containing only ddBTP (bottom). The total number of ssDNA calls and total ssDNA call burden (calls per base) are shown. **g**, Fraction of ssDNA calls that are T > A (corrected for trinucleotide context opportunities) versus the ssDNA T > A burden in post-mortem liver (n = 5) and kidney (n = 5) samples profiled by HiDEF-seq without A-tailing (i.e., Klenow reaction -dATP/+ddBTP). **h**, Concordant dsDNA mutation burdens in sperm sample SPM-1013 measured by HiDEF-seq with A-tailing (i.e., Klenow reaction +dATP/+ddBTP) and without A-tailing (i.e., Klenow reaction -dATP/+ddBTP). yo, years old. **i**, Mutational signature contribution to dsDNA mutations detected by HiDEF-seq in primary human tissues from individuals without cancer predisposition. Post-mortem liver and kidney samples were profiled by HiDEF-seq without A-tailing. All samples, except blood from a 62-year-old individual with a history of kidney disease (1901, asterisk), were jointly analysed with fitting of SBS1 and de novo extraction of one additional signature SBSiii. Blood samples of the 62-year-old profiled by HiDEF-seq were analysed separately (plot shows average signature contributions across 5 blood samples) due to identification of an additional signature SBSiv. Analysis of samples grouped by tissue type, excluding the 62-year-old blood sample, produced similar results. For de novo extracted signatures (SBSiii and SBSiv), the cosine similarities to the most similar COSMIC signatures are shown in parentheses. Sperm, kidney and liver samples from an infant (1443) and 18-year-old (1409), and blood from a 4-year-old (5203) are not included here since their number of mutations are too low for reliable signature extraction. **e**,**h**, Bars (e) and dots (h) show point estimates, and error bars are their Poisson 95% confidence intervals. **e**-**g**, Rxn, reaction.

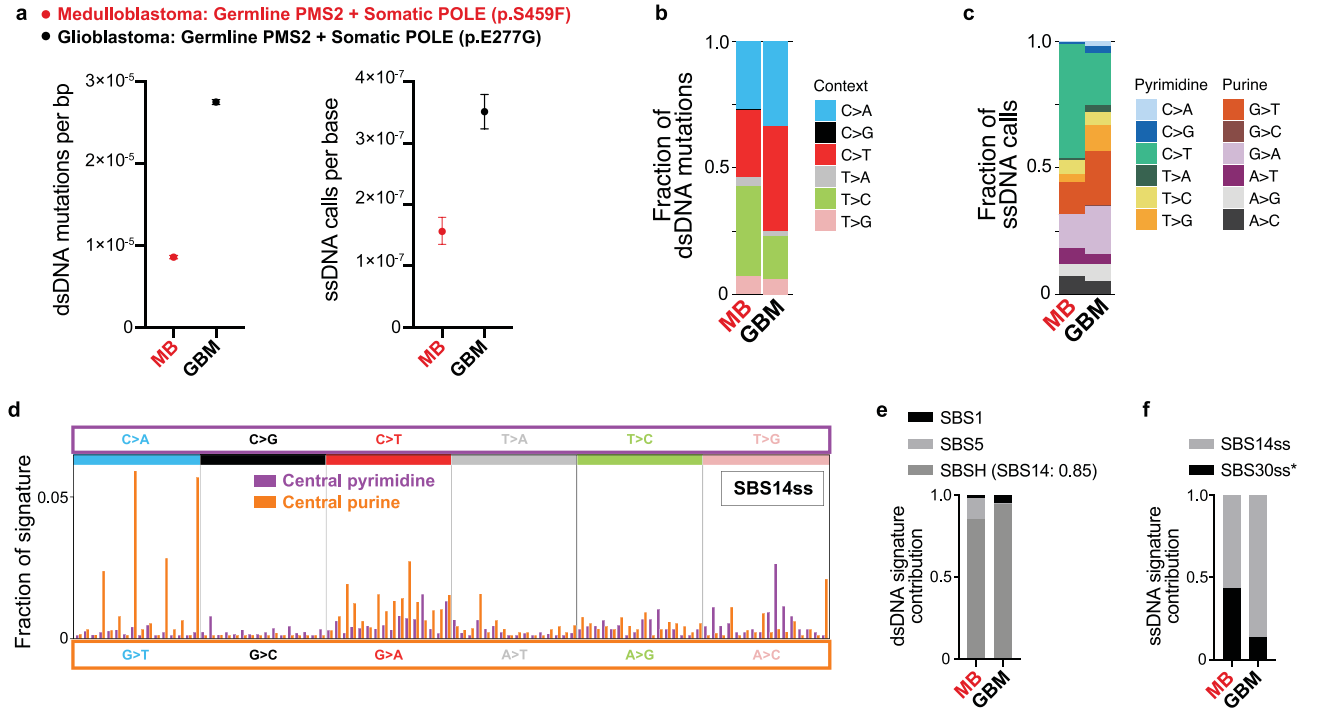


Extended Data Fig. 6 | Comparison of HiDEF-seq and NanoSeq. a, Comparison of HiDEF-seq versus NanoSeq dsDNA mutation spectra for individual 63143. **b**, Comparison of HiDEF-seq versus NanoSeq ssDNA call burdens, separated by call type. For each call type (i.e., C > A, C > G, etc.), each bar represents a different sample. Samples for each call type, from left to right, are 1105 and 6501 for healthy blood; 63143 for POLE blood; and 1443 for kidney. Comparison for sperm samples is shown in Fig. 1g. **c**, Comparison of HiDEF-seq versus NanoSeq ssDNA call spectra for 6501 (Blood, 43 yo), 63143 (POLE blood), and SPM-1060 (sperm, 49 yo). **a**-**c**, yo, years old; mo, months old.

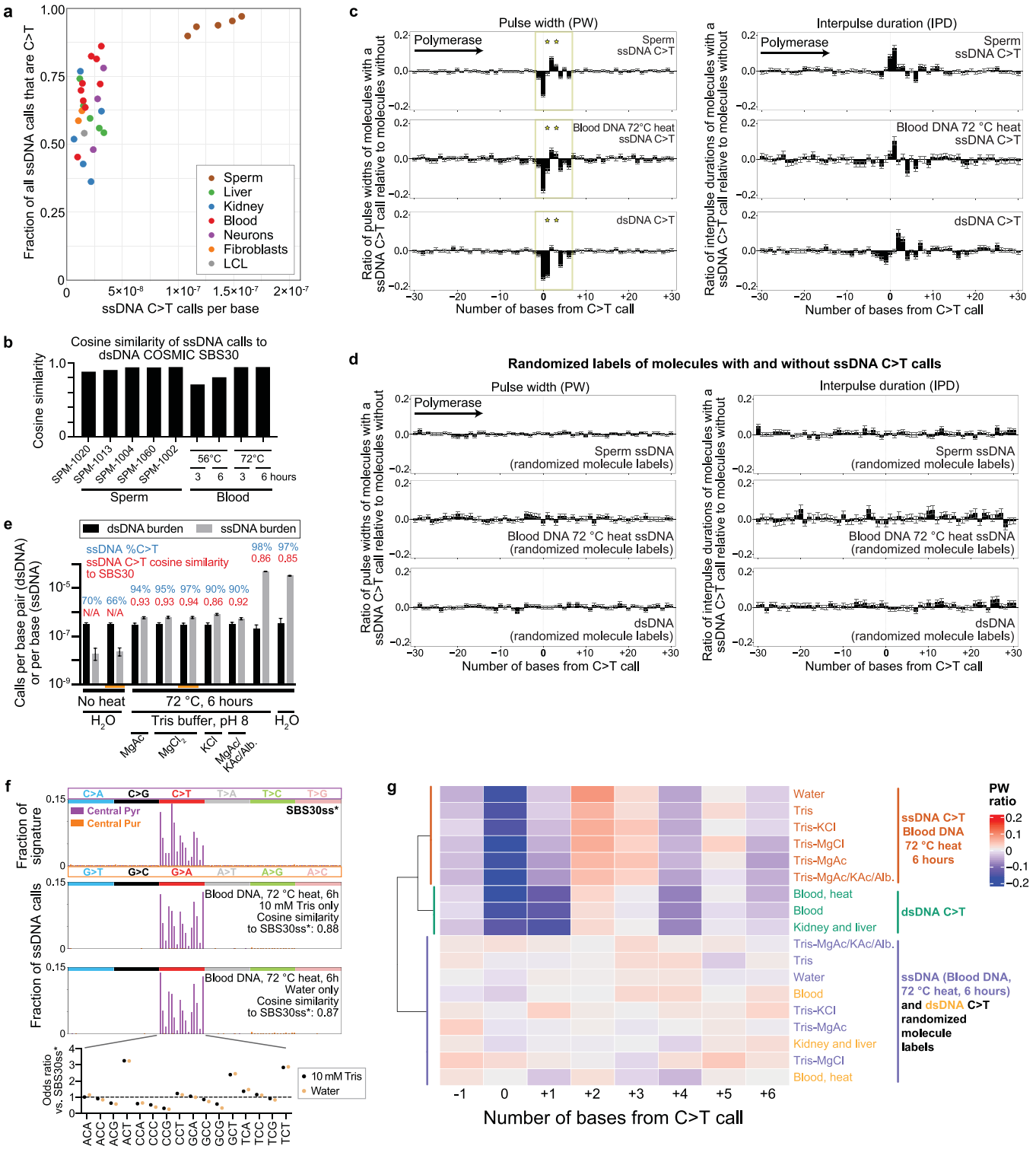


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | dsDNA mutation burdens and patterns in cancer predisposition syndromes. a, Fraction of dsDNA mutations in each context. Non-cancer predisposition samples are (left to right): Blood (B) 5203, 1105, 1301, 6501, and 1901; lymphoblastoid cell line (LCL) GM12812; primary fibroblasts GM02036 and GM03348. Cancer predisposition samples are (left-to-right, in the same order and annotated sample types as top-to-bottom cancer predisposition samples in panel (c)): GM16381, GM01629, GM28257, 55838, 58801, 57627, 1400, 1324, 1325, 60603, 59637, 57615, 63143 (LCL), 63143 (B), CC-346-253, CC-388-290, CC-713-555. Affected genes annotated below. Note, GM02036 (asterisk) has a significant increase in C > T mutations with a spectrum matching COSMIC SBS7a (ultraviolet light exposure), likely due to the fibroblasts deriving from sun-exposed skin. **b**, Representative dsDNA mutation spectra of a sample for each affected gene, corrected for trinucleotide context opportunities. Sample IDs are in parentheses. Ages (yo, years old) are listed for blood samples. **c**, Fraction of dsDNA mutations attributable to de novo extracted dsDNA mutational signatures. Sample genotypes are on the right (hom., homozygous; compound heterozygous variants separated by '/'). In parentheses is the cosine similarity to the most similar COSMIC signature when the similarity is ≥ 0.8 (weak similarity: 0.8 – 0.85; moderate similarity: 0.85 – 0.9; strong similarity: ≥ 0.9; Methods). In *ERCC6* and *ERCC8* mutant cell lines, whose mutational patterns are unknown, we identified signature SBSB with weak similarity (cosine similarity 0.82) to the COSMIC SBS36 signature. For SBSF, the most similar COSMIC signature is SBS10c, but the cosine similarity of 0.79 is not considered significant. For SBSG, the most similar COSMIC signature is SBS40, but the cosine similarity of 0.76 is not considered significant. SBSG had non-significant similarities to SBS18 (0.69) and SBS36 (0.59), which have been previously associated with *MUTYH*[21]. These *MUTYH* signatures were not extracted due to the normal mutation burdens of our *MUTYH* blood samples (see panel (d)), which is expected at these sample ages and our interrogated base coverage[21]. Note that SBS40 resembles SBS18 and SBS36 in the C > A spectrum that is enriched in *MUTYH* syndrome[21]. Signature extraction was performed for samples of each DNA repair pathway (except *XPC* separately from *ERCC6*/*ERCC8*), while simultaneously fitting COSMIC SBS1 and SBS5 (Methods). Samples are in the same top-to-bottom order as left-to-right cancer predisposition samples in panel (a). **d**, dsDNA mutation burden per base pair divided by the age of the individual in years at the time of blood collection, corrected for trinucleotide context opportunities and sensitivity. Only blood samples are shown, since blood can be annotated with the age of the individual. Accordingly, since we did not profile blood samples nucleotide excision repair syndrome, this category is not shown. Non-cancer predisposition blood samples are the same (left-to-right) as in panel (a) (left-to-right). Cancer predisposition blood samples are the same (left-to-right) as blood samples in panel (c) (top-to-bottom). Affected genes annotated below. **e**, Replication strand asymmetry based on replication timing data (Methods) of AGA > ATA ssDNA mismatches and dsDNA mutations in *POLE* PPAP samples. Reference (+) refers to the human reference genome plus strand. Non-reference (-) strand lagging and leading strand synthesis corresponds to negative and positive fork polarity values, respectively (Methods). The 'strand ratio' (Y-axis) is calculated as the fraction of all AGA > ATA non-reference strand events that have the specified fork polarity divided by the fraction of all AGA > ATA reference strand mutations that have the specified fork polarity (Methods). *, $p = 0.015$; ***, $p < 10^{-15}$ (chi-squared test; n = 73 ssDNA AGA > ATA mismatches; n = 3,871 dsDNA AGA > ATA mutations). For dsDNA mutations, bars show the average across PPAP samples (n = 4), and for ssDNA mismatches, due to their low number, bars show a single estimate for calls pooled across PPAP samples. See (f) for analysis of dsDNA mutations separated by fork polarity quantiles (rather than positive versus negative polarity), which cannot be plotted for ssDNA mismatches due to the low number of ssDNA mismatches per quantile. ssDNA strand ratios were calculated using calls of all *POLE* PPAP samples, since there are too few calls to reliably analyse individual samples. dsDNA strand ratios were calculated separately for each sample (plot shows average and standard deviation). Excluding calls overlapping genes to exclude transcription strand biases was still significant for dsDNA mutations ($p < 10^{-15}$) but not ssDNA mismatches, but the latter had significantly reduced power due to a 55% reduction in the number of analysed ssDNA calls. **f**, Replication strand asymmetry of AGA > ATA dsDNA mutations in *POLE* PPAP samples calculated for each fork polarity quantile. Fork polarity quantiles divide fork polarity values into 9 quantile bins from 0 to 1, with higher values corresponding to a greater probability of the non-reference strand being replicated in the leading rather than lagging strand direction (Methods). Random loci are the average of 50 sets of 1,000 random genomic loci with either the sequence AGA or TCT for which there is replication timing data at the locus. The 'strand ratio' is calculated for *POLE* PPAP samples as in (e), and it is calculated for random genomic loci as the fraction of all AGA non-reference strand loci that are in the fork polarity quantile bin divided by the fraction of all AGA reference strand loci that are in the fork polarity quantile bin. PPAP samples are the same top-to-bottom order in the legend as top-to-bottom PPAP samples in (c). Asterisks signify statistical significance in comparison of the *POLE* PPAP 4-sample average (dashed line) to random loci (heteroscedastic two-tailed t.test); p-values left-to-right for asterisks: $3.7 \cdot 10^{-17}$, 0.001, 0.009, 0.02, 0.003. Excluding mutations overlapping genes to exclude transcription strand biases produced similar results ($p = 3.1 \cdot 10^{-10}$, 0.003, and 0.04 for quantiles 0-0.1, 0.1-0.2, and 0.6-0.7, respectively), but this analysis has reduced power due to the 55% reduction in the number of mutations. **a-f**, See additional samples details in Supplementary Tables 1–4. **e,f**, Error bars: standard deviation.

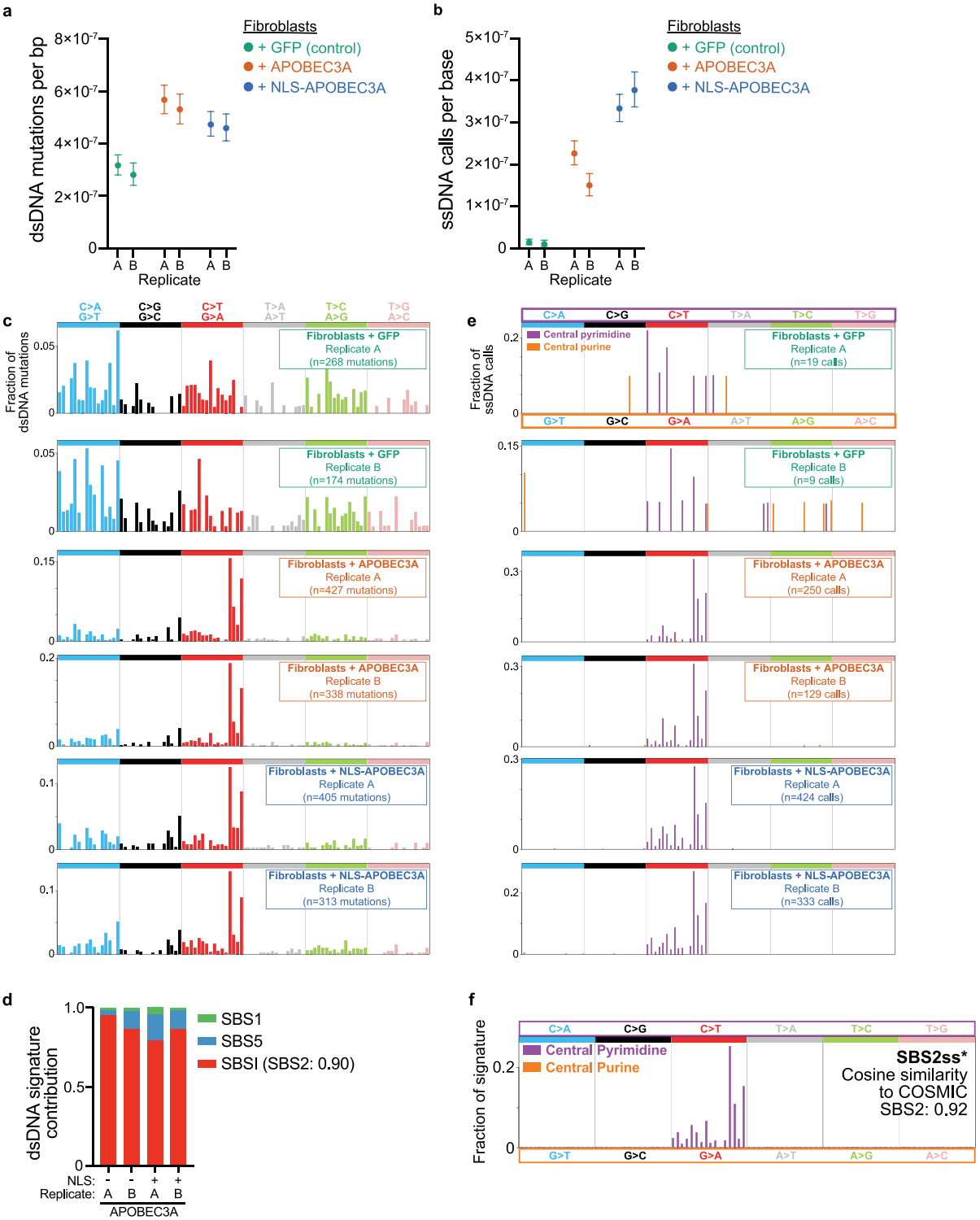


Extended Data Fig. 8 | Hypermutating tumours deficient in both mismatch repair and polymerase proofreading. a, Burdens of dsDNA mutations (left) and ssDNA calls (right). Burdens are corrected for trinucleotide context opportunities and detection sensitivity (Methods). **b,c**, Fraction of dsDNA mutation burdens (b) and ssDNA call burdens (c) by context, corrected for trinucleotide context opportunities. **d**, ssDNA mismatch signature SBS14ss extracted from tumour samples, while simultaneously fitting SBS30ss*. **e**, Fraction of dsDNA mutations attributed to each dsDNA signature. Cosine similarity of the extracted signature SBSH to the most similar COSMIC SBS signature is shown in parentheses. Cosine similarities of original spectra of samples to spectra reconstructed from component signatures are (left to right): 0.94 and 0.998. **f**, Fraction of ssDNA calls attributed to each ssDNA signature. Cosine similarities of original spectra of samples to spectra reconstructed from component signatures are (left to right): 0.91 and 0.98. **a**, Dots and error bars: point estimates and their Poisson 95% confidence intervals. **a-c,e,f**, MB, medulloblastoma (ID: Tumour 8); GBM, glioblastoma (ID: Tumour 10). See Supplementary Table 1 for sample details.

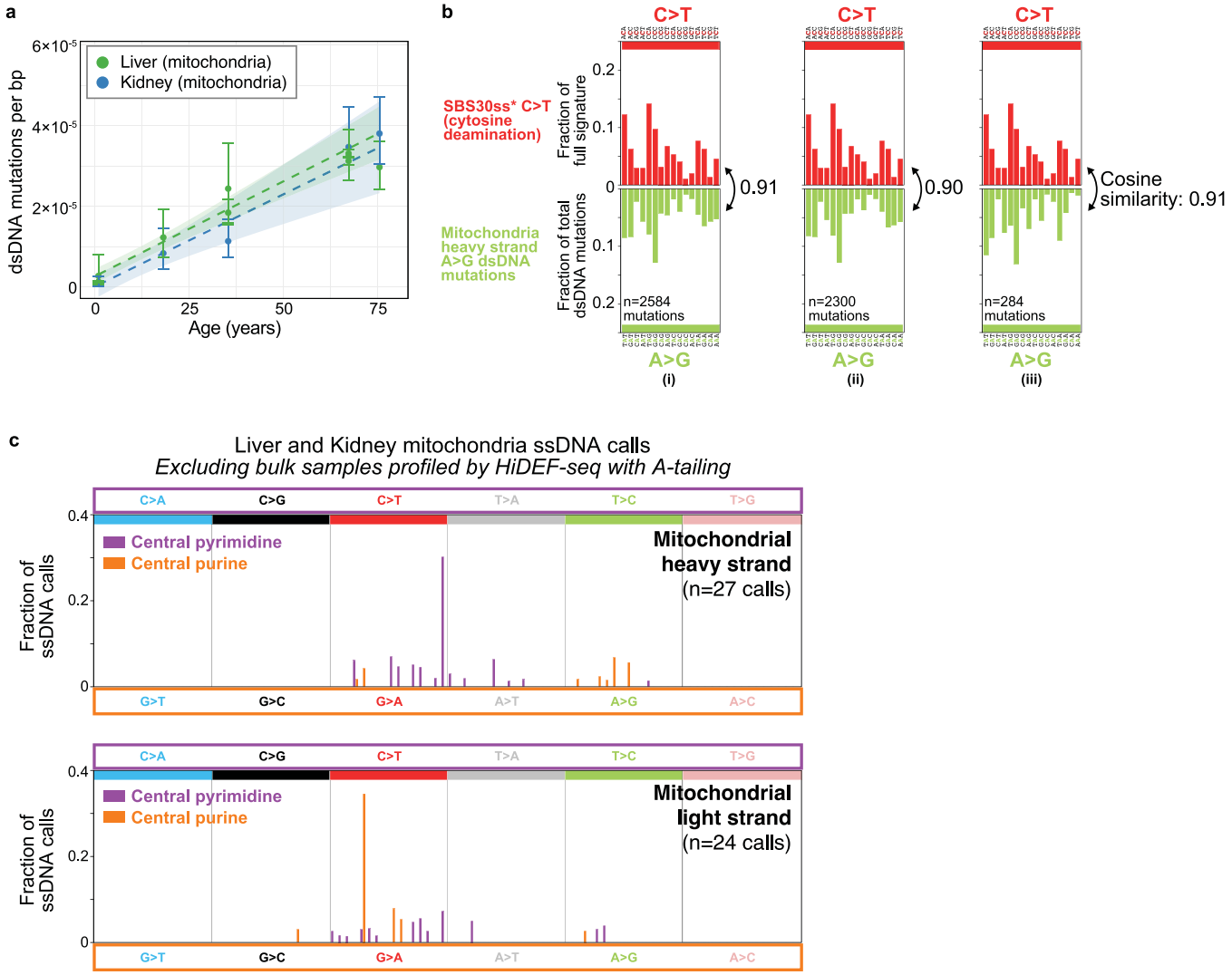


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Burdens of ssDNA C > T calls, kinetic interpulse duration profiles, and profiling of heat treatment in varied buffers. **a**, Fraction of ssDNA calls that are C > T (corrected for trinucleotide context opportunities) across all HiDEF-seq samples from healthy individuals and cell lines (i.e., excluding cancer predisposition syndromes), versus the ssDNA C > T burden. Data shown for liver and kidney samples profiled by HiDEF-seq without A-tailing. Sperm consistently have the highest fraction of ssDNA calls that are C > T. LCL, lymphoblastoid cell line. **b**, Cosine similarity of ssDNA call spectra to SBS30 after projecting ssDNA spectra to central pyrimidine contexts. **c**, Average ratio of pulse widths (left) and interpulse durations (right) at C > T calls and 30 flanking bases relative to molecules aligning to the same locus without the call (sperm: n = 1799 calls; blood DNA 72 °C heat, 3 and 6 h: n = 626 calls; dsDNA C > T mutations in a larger set of non-heat treated blood DNA, 56 °C and 72 °C heat treated blood DNA, sperm, kidney, and liver samples: n = 1202 mutations; Methods). Positions +1 and +3 (stars) best discriminate ssDNA C > T damage from dsDNA C > T mutations. Yellow box is the span shown in Fig. 4f. **d**, Average ratio of pulse width (left column) and interpulse duration (right column) after randomizing labels of molecules with and without the calls, for the same samples and calls as in panel (c). **e**, dsDNA mutation and ssDNA call burdens of heat-treated blood DNA in an additional experiment testing the effect of different buffers and different DNA extraction methods (orange underline, Puregene alcohol precipitation; all other samples, MagAttract with magnetic beads). MgAc, magnesium acetate; $MgCl_2$, magnesium chloride; KCl, potassium chloride; KAc, potassium acetate; Alb, albumin; Tris buffer is Tris-HCl except for the MgAc/KAc/Alb that is Tris-Acetate (see Supplementary Table 1 for concentrations). Non-heat treated DNA samples were placed on ice for 6 h. The percentage of ssDNA sequencing calls that are C > T are annotated above each sample. Cosine similarity to COSMIC dsDNA signature SBS30 is annotated below each sample, after collapsing ssDNA calls to central pyrimidine trinucleotide contexts and correcting for trinucleotide context opportunities, except for the no-heat treatment samples that do not have sufficient C > T calls ('N/A'). **f**, SBS30ss* signature (reproduced from Fig. 4d) compared to spectra of ssDNA calls after 72 °C heat damage of blood DNA for 6 h (h) in only 10 mM Tris buffer (n = 10,852 calls) or only water (n = 2,751 calls). Spectra are plotted after correcting for trinucleotide context opportunities. Bottom, odds ratios of spectrum contributions at C > T contexts of the Tris-only and water-only samples compared to SBS30ss* (which was derived from sperm and salt-buffer heat-treated samples). Pyr, pyrimidine, Pur, purine. **g**, Heat map of average pulse width ratios for ssDNA and dsDNA C > T calls for positions −1 to +6, for blood DNA samples heated at 72 °C for 6 h in different buffers or water, and for additional samples for comparison. Unbiased clustering (dendrogram) separates kinetic profiles of ssDNA C > T calls from dsDNA C > T calls and from kinetic profiles after randomizing labels of molecules with and without the calls. dsDNA 'Blood, heat': blood DNA heat-treated at 56 °C and 72 °C (both 3 h and 6 h for each); dsDNA 'Blood': 4 samples, not heat treated. dsDNA 'Kidney and liver': 10 samples, not heat treated. **b**, HiDEF-seq spectra are corrected for trinucleotide context opportunities. **c,d**, Error bars: standard error of the mean. **e**, Bars and error bars: point estimates and their Poisson 95% confidence intervals.



Extended Data Fig. 10 | APOBEC3A-induced dsDNA and ssDNA call burdens and patterns. a,b, Burdens (corrected for trinucleotide context opportunities and sensitivity) of dsDNA mutations (a) and ssDNA calls (b) in fibroblasts transduced with lentivirus-expressing green fluorescent protein (GFP) as a control or APOBEC3A with or without a nuclear localization signal (NLS). Two biological replicates are shown for each condition. **c**, Spectra of dsDNA mutations corrected for trinucleotide context opportunities. **d**, Fraction of dsDNA mutations attributed to each dsDNA signature. Cosine similarity of the de novo extracted signature SBSI to the most similar COSMIC SBS signature is shown in parentheses. Cosine similarities of original spectra of samples to spectra reconstructed from component signatures are (left to right): 0.99, 0.98, 0.98, and 0.97. **e**, Spectra of ssDNA calls corrected for trinucleotide context opportunities. **f**, SBS2ss* obtained by de novo signature extraction from APOBEC3A samples. Cosine similarity to SBS2 is calculated after projecting to central pyrimidine trinucleotide context. **a,b**, Error bars: Poisson 95% confidence intervals.



Extended Data Fig. 11 | Mitochondrial genome dsDNA mutation rates, similarity between SBS30ss* and mitochondrial genome heavy strand A > G dsDNA mutations, and mitochondrial ssDNA call spectra. a, Mitochondrial dsDNA mutation burdens versus age in liver and kidney samples, including liver samples from which mitochondria were enriched. Dashed lines: weighted least-squares linear regression. Shaded ribbon: 95% confidence interval. **b**, SBS30ss* (cytosine deamination) spectrum is projected to central pyrimidine trinucleotide contexts and compared to mitochondria heavy strand A > G dsDNA mutation spectrum (corrected for trinucleotide context opportunities), for different sample sets: (i) HiDEF-seq liver and kidney samples, including liver samples from which mitochondria were enriched (i.e., same set of samples in Fig. 5a, c and Extended Data Fig. 11a); (ii) 5697 purified liver mitochondria samples only (plot includes 89% of the mutations in (i)); (iii) Sample set (i), excluding the 5697 purified liver mitochondria samples (plot includes 11% of the mutations in (i)). Note, the contexts of SBS30ss* are matched with the reverse complement flanking base contexts of mitochondria heavy strand A > G mutations. The number of dsDNA A > G mutations is indicated. **c**, Spectrum of mitochondrial ssDNA calls combined from the liver and kidney samples shown in Fig. 5a, c and Extended Data Fig. 11a. The spectrum is corrected for trinucleotide context opportunities, separately for each strand. See Fig. 5d for a spectrum that includes bulk (i.e., non-mitochondria enriched) samples profiled by HiDEF-seq with A-tailing. **a**, Dots and error bars: point estimates and their Poisson 95% confidence intervals.

nature portfolio

Corresponding author(s): Gilad D. Evrony

Last updated by author(s): May 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	Software tools: BWA-MEM v0.7.17, GATK v4.1.9.0, DeepVariant v1.2.0, pbccs v5.0.0 (ccs, Pacific Biosciences), pbmm2 v1.7.0, lima v2.5.0 (Pacific Biosciences), HiDEF-seq v1.1 (https://github.com/evronylab/HiDEF-seq), R v4.1.2, bcftools v1.14, samtools v1.14, wigToBigWig v2.8, wiggletools v1.2.11, zmwfilter v1.2.0 (Pacific Biosciences), SeqKit v2.1.0, KMC v3.1.1 (https://github.com/refresh-bio/KMC), NanoSeq v3.2.1 (https://github.com/cancerit/NanoSeq), REAPR v1.0.1834, SMALT v0.7.6. R packages: GenomicAlignments v1.30.0, GenomicRanges v1.46.1, vcfR v1.12.0, Rsamtools v2.10.0, plyr v1.8.6, configr v0.3.5, MutationalPatterns v3.4.1, magrittr v2.0.2, readr v2.1.2, dplyr v1.0.8, plyranges v1.14.0, stringr v1.4.0, digest v0.6.29, rtracklayer v1.54.0, qs v0.25.2, sigfit v2.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data generated in this study (FASTQ files for Illumina sequencing; subreads BAM files for PacBio data) are available at the NCBI database of Genotypes and Phenotypes with accession phs003604 (all samples except those from the International Replication Repair Deficiency Consortium and subjects D1 and D2) and at the European Genome-phenome Archive with accession EGAS50000000318 (samples from the International Replication Repair Deficiency Consortium). Sequencing data of subjects D1 and D2 is not deposited in these databases due to consent limitations. See Supplementary Table 1 for accession IDs of specific samples.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex is recorded for all samples included in the study in Supplementary Table 1. Sex differences in mutation and single-strand call burdens were not assessed, as this was not a primary outcome of the study and the study lacks statistical power to investigate this question.
Population characteristics	Sex, age, genotype, and disease details are provided for all samples in Supplementary Table 1.
Recruitment	Participants were recruited through several IRB/ethics board-approved human subjects protocols. Recruitment criteria were as follows: 1) NYU protocol 1: individuals from families with rare genetic diseases. 2) NYU protocol 2: individuals who are healthy sperm donors. 3) Hospital for Sick Children: individuals with rare DNA repair syndromes. 4) Cryos: individuals who are healthy sperm donors. 5) U. Pittsburgh: Individuals with cancer predisposition syndromes. There are no anticipated self-selection biases beyond the inclusion criteria of each study that would affect the results of our study, because there are no feasible such confounders that would correlate with the mutational processes we are studying.
Ethics oversight	All samples were collected under human subjects research protocols approved by one of: New York University Grossman School of Medicine Institutional Review Board The Hospital for Sick Children Research Ethics Board Cryos International Sperm Bank scientific advisory committee University of Pittsburgh Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for healthy tissues was determined based on power calculations to quantify double-strand mutation rates with regression 95% confidence intervals within 20% of the estimated mutation rate. Sample size for rare disease samples (i.e. cancer-predisposition syndromes) was determined based on sample availability.
Data exclusions	No data was excluded from the analysis.
Replication	Eight of the samples had 2 technical replicates each, with concordant results. See Supplementary Table 2 for data.
Randomization	Randomization was not performed, because there was no intervention to randomize.
Blinding	Blinding was not performed, because there was no observation or subject interaction that would be susceptible to observer bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

NeuN-Alexa-647: abcam product # ab190565
TOM22: Miltenyi Biotec, within Mitochondria Isolation Kit, human, product #130-094-532

Validation

NeuN antibody was validated by the manufacturer in human and rodent brain tissue. We also observe the expected distribution of nuclei populations in flow cytometry.
TOM22 antibody was validated by Miltenyi Biotec by confirming purification of mitochondrial proteins from human mitochondria isolated using the antibody. Specifically, after mitochondria isolation using this antibody, the manufacturer performed a Western blot confirming presence of the COX1 mitochondrial protein and absence of the KDEL endoplasmic reticulum protein (see https://www.miltenyibiotec.com/US-en/products/mitochondria-isolation-kit-human.html#130-094-532).

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

ID,Sex,Line type,Source (LCL = Lymphoblastoid Cell Line; SickKids = Hospital for Sick Children)
57627 Male LCL SickKids
60603 Male LCL SickKids
GM12812 Male LCL Coriell
58801 Male LCL SickKids
55838 Female LCL SickKids
57615 Female LCL SickKids
59637 Male LCL SickKids
GM02036 Female Primary fibroblasts Coriell
GM03348 Male Primary fibroblasts Coriell
GM16381 Male Primary fibroblasts Coriell
63143 Female LCL SickKids
GM01629 Female Primary fibroblasts Coriell
GM28257 Male Primary fibroblasts Coriell
Lenti-X 293T cells from Takara

Authentication

Cell lines from Coriell were authenticated by Coriell by human identity microsatellite genotyping. Cell lines from SickKids were authenticated by confirming in germline sequencing data performed in this study the presence of the correct pathogenic mutation in the relevant cancer-predisposition gene. The cell line from Takara was authenticated by Takara based on STR markers and morphology.

Mycoplasma contamination

Cell lines from Coriell were tested for mycoplasma contamination by Coriell and found to be negative. Cell lines from SickKids were not tested for mycoplasma contamination. The cell line from Takara was tested for mycoplasma contamination by Takara and found to be negative.

Commonly misidentified lines
(See ICLAC register)

None

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Nuclei were isolated from post-mortem brain tissue by dounce homogenization and sucrose gradient centrifugation as described in the Methods.
Instrument	SONY LE-SH800
Software	SONY SH800S software
Cell population abundance	Sorted NeuN+ nuclei were reanalyzed on the same flow sorting instrument to confirm > 99% purity.
Gating strategy	1. Scatter gate to remove debris: BSC-A vs FSC-A 2. Doublet gate: FSC-H vs FSC-A 3. NeuN+ nuclei: NeuN-647-A vs FSC-A (See Supplementary Note 12 for representative figure)

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.