Eukaryotic cells have many strategies to ensure the integrity of their genomes, with high-fidelity DNA replication and DNA-repair processes countering exogenous and endogenous DNA lesions. The process of transcription targets DNA repair machinery to expressed genes, preferentially reducing their mutation rate after DNA damage. Despite this targeted repair, in microorganisms, the process of transcription itself is mutagenic—a phenomenon that is referred to as transcription-associated mutagenesis (TAM). In yeast, topoisomerase 1 (TOP1) activity is a major source of TAM and results in a distinctive transcription-dependent signature of 2–5 bp deletions at tandem repeat sequences. Genome-embedded ribonucleotides have been established as a cause of TAM deletions in yeast. Such ribonucleotides are frequently incorporated by DNA polymerases during replication, and represent the most prevalent aberrant nucleotides in the eukaryotic genome. These genome-embedded ribonucleotides are normally removed by ribonucleotide excision repair (RER), a process initiated by the heterotrimeric ribonuclease H2 enzyme. However, when Top1 cleaves at embedded ribonucleotides instead of RNase H2, this can result in small deletions.

In the last decade, the widespread use of genome sequencing has enabled unbiased sampling of human mutations, substantially advancing understanding of mutagenesis in the germline and in neoplasia. Multiple mutational processes act during cancer evolution, and mathematical methods decomposing tumour mutational profiles have been developed to define signatures that may correspond to individual mutagenic mechanisms. This has successfully defined cell-intrinsic, environmental and treatment-related origins for many base-substitution signatures in cancer. However, the origin of a substantial number of signatures remains unknown, and some may be artefactual. Recently, cancer signature analysis has been extended to indels, small (1–49 bp) insertions and deletions. Such indels are an important class of mutations that contribute substantially to disease-causing germline variants (>20%) and human variation.

Here we investigate an indel signature of unknown cause—ID4. We show experimentally that ID4 deletions are increased in RNase-H2-deficient cell lines and cancers and delineate a human TOP1-mediated TAM signature (ID-TOP1) that is relevant to both somatic and germline mutagenesis.
ID4, a distinct cancer indel signature

The ID4 cancer signature, as categorized by COSMIC, comprises small deletions (typically 2–5 bp in size) of one repeat unit at SSTR and MH sites. Repeated sequences (i–vi) are shown in bold and colour. Deletions are shown in red. b, Indel mutations similar to those detected in ID4 accumulate genome-wide in yeast with high levels of genome-embedded ribonucleotides. Reanalysis of WGS data for *mh201Δ* pol2-M644G yeast. c, Schematic of a frameshift mutation reporter containing many 2 bp SSTRs. Frameshift mutations in HygroR result in neomycin-resistant yeast colonies. P (i), TEF promoter; P2A, self-cleaving peptide. d, e, Fluctuation assays demonstrated that Top1-mediated 2 bp SSTR mutations occur in wild-type and RNase-H2 deficient (*rnh201Δ*) backgrounds. d, Mutation rates for n = 16 independent cultures per strain. Data are median ± 95% confidence intervals. e, WT and *mh201Δ* have similar indel mutation spectra, and differ from top1Δ strains. Spectra of neomycin-resistant colonies. n indicates the number of independent indels detected. Cosine similarity P values were empirically determined (Extended Data Fig. 2e, f). Del, deletion; ins, insertion.

ID4 resembles a yeast mutation signature

Noting similarities to a Top1-induced TAM (Top1-TAM) in *Saccharomyces cerevisiae*, we reanalysed published genome-wide mutation-accumulation experiments performed with *mh201Δ* pol2-M644G yeast. This strain is particularly susceptible to Top1-TAM as it accumulates genome-embedded ribonucleotides at high levels due to RNase H2/RER deficiency and enhanced ribonucleotide incorporation by a steric-gate mutation at the catalytic site of the replicative polymerase Polε. Similarities to the ID4 signature were apparent with a comparable pattern of small deletions at SSTRs, although mutational events at sites of SNMH were not evident in the yeast data. As more than 1 million ribonucleotides are incorporated by DNA polymerases per replicating mouse cell, we reasoned that genome-embedded ribonucleotides might cause similar mutational events in mammalian cells. To experimentally assess whether TAM contributes to indel formation in human RER-deficient cells, we developed a reporter to enable sensitive and specific detection of mutational events arising from Top1 activity in both yeast and mammals.

**Top1-dependent deletions in yeast**

Mutation rates are routinely measured in *S. cerevisiae* using well-characterized but species-specific selectable markers (LYS2, URA3, CAN1). Thus, to establish a system that could be transferred between yeast and mammalian cells, we used an approach inspired by the Traffic Light reporter assay, incorporating both positive and negative selection cassettes in a single transcriptional unit (Fig. 1C). The hygromycin-resistance gene (HygroR) was used both as the
mutational target and negative selection marker. Indels causing a 2 bp frameshift within HygroR, including 2 bp deletions, result in translation of an otherwise out-of-frame P2A self-cleaving peptide and the neomycin-resistance (NeoR) gene, permitting positive selection of mutated colonies with neomycin (Extended Data Fig. 2a). To enrich the target for 2 bp tandem repeats, in silico redesign incorporated synonymous substitutions such that SSTRs accounted for >50% of the HygroR open reading frame.

For validation, the reporter was inserted into the S. cerevisiae genome and fluctuation assays were performed to assess the mutation rates in strains deficient for RER and/or Top1. A 37-fold increase in mutation rate was observed for the rnh201Δ top1Δ double-mutant strain, consistent with Top1-dependent mutagenesis at genome-embedded ribonucleotides. Notably, there was a 10-fold decrease in the mutation rate for top1Δ compared with the wild-type strain, and a 35-fold decrease in 2 bp SSTR deletions (Extended Data Fig. 2b), consistent with previous reports. Furthermore, the observed mutational spectrum was most similar for wild-type and mnh201Δ strains, but substantially different compared with the top1Δ and mnh201Δ top1Δ strains (Fig. 1e and Extended Data Fig. 2c–f). Taken together, we conclude that the same Top1-mediated mutations occur, albeit at different frequencies, in wild-type cells when RER is functional and in RNase-H2-deficient strains when elevated levels of ribonucleotides are present in the genome.

**Fig. 2** | SSTR deletions of 2 bp are increased in RNase-H2-null HeLa cells. a | Schematic of the reporter targeting the AAVS1 safe harbour locus to generate reporter cells (Extended Data Fig. 3). HA, homology arm; L, left; R, right. b, c | Validation of RNASEH2A-KO reporter clones. b | Immunoblot analysis of cell lysates detecting the three RNase H2 subunits. GAPDH was used as the loading control. Gel source data are provided in Supplementary Fig. 1. c | Cellular RNase H2 enzyme activity. Data are mean ± s.d. n = 3 technical replicates. HeLa, no modification; parental, HeLa with reporter (grey); KO1 and KO2, CRISPR-mediated RNASEH2A-KO clones (red); RNASEH2A+, CRISPR-edited reporter clone retaining RNase H2 activity (green). d | Fluctuation assays establish a significantly increased mutation rate in RNase-H2-null (KO) cells (P = 2 × 10−4). Statistical analysis was performed using a two-sided Mann-Whitney test. Data are median ± 95% confidence intervals. The data points show the rates for independent cultures. n = 9 (RNase H2 proficient, RNASEH2A+); n = 10 (K01, open circles) and n = 6 (K02, open squares). e | 2 bp SSTR and SNMH deletions are frequent in both RNASEH2A+ and RNASEH2A-KO cells. Indel mutation spectra. n shows the number of indels identified by sequencing colonies from independent cultures.

**TOP1-mediated mutations in human cells**

Having validated the reporter in yeast, the same 2 bp repeat-enriched HygroR sequence was used to determine whether TOP1-mediated mutagenesis at embedded ribonucleotides is conserved in human cells (Fig. 2 and Extended Data Fig. 2g).NeoR was replaced by the puromycin-resistance (PuroR) gene, with reporter expression driven from the mammalian ubiquitous CAG promoter, permitting rapid antibiotic selection in mammalian cells. This modified reporter was inserted at the AAVS1 safe harbour locus in HeLa cells (Fig. 2a and Extended Data Fig. 3a–e). CRISPR-Cas9-mediated genome editing targeting the catalytic site of RNASEH2A was then used to generate two independent knockout (KO) reporter clones, alongside a control clone that had also been processed through the editing and clonal selection steps (Fig. 2b, c and Extended Data Fig. 3). The control clone retained RNase H2 activity, whereas there was complete loss of cellular RNase H2 activity in the KO clones, accompanied by high levels of ribonucleotides in genomic DNA (Fig. 2b, c and Extended Data Fig. 3f, g).

In fluctuation assays, RNase-H2-null clones demonstrated a significant 3.1-fold increase in mutation rate (Fig. 2d) and 5.2-fold more 2 bp SSTR deletions (Extended Data Fig. 3h) compared with RNase-H2-proficient cells (RNASEH2A+), consistent with conservation of TOP1-directed mutagenesis in human cells. As in yeast (Fig. 1e), the overall mutational profile of reporter mutations was similar between RNase-H2-proficient and null HeLa cells (cosine similarity = 0.89, P < 10−4), predominantly comprising 2 bp SSTR deletions (Fig. 2e).

The mutation rate for RNase-H2-null HeLa cells (8.0 × 10−9 per bp per generation; 95% confidence interval = 6.7–9.5 × 10−9) was similar to that observed for mnh201Δ yeast (Fig. 1d), whereas the rate was substantially higher for RNASEH2A+ control cells compared with wild-type yeast. However, the increased mutation rate in RNase-H2-null HeLa cells probably underestimates the true impact of RER deficiency in human cells as, despite the fact that the control RNASEH2A+ HeLa reporter cells retained protein expression (Fig. 2b), the clone also had acquired mutations at the CRISPR editing site that reduced enzymatic activity (Fig. 2c), causing a moderate increase in genomic ribonucleotide content (Extended Data Fig. 3f, g).

To confirm these findings, we used a complementary approach to establish the relevance of such mutational events genome-wide, performing mutation-accumulation experiments using human hTERT RPE-1 (TP53+) diploid cell lines. Ancestral populations for RNase-H2-wild-type and RNase-H2-null cells (RNASEH2A-KO or RNASEH2B-KO; Extended Data Fig. 4a–d) were established after initial
Article

Fig. 3 | ID4 SSTR and MH mutations are increased genome-wide in RNase-H2-deficient RPE-1 cells. a, Schematic of the mutation-accumulation experiment. Long-term culture of hTERT-RPE-1 TP53+/− RNase-H2-wildtype (WT) and RNase-H2-null cell lines (RNASEH2A KO (AKO), RNASEH2B KO (BKO)) bottlenecked every 25 doublings by single-cell sorting. b, Mutations acquired during long-term culture were significantly enriched for 2–5 bp deletions in RNase-H2-null cells, but the other mutation categories were not (Extended Data Fig. 4e). Data are mean ± s.d. Statistical analysis was performed using two-sided Fisher’s exact tests with Bonferroni correction, comparing wild type (counts pooled from n = 3 independent clones) versus KO (n = 2 independent clones) for 2–5 bp deletions versus all of the other indel types. c, d, ID4 occurs in RNase-H2-null cells (c) and is the major signature after subtracting background mutations that are observed in wild-type cells (d).

MH deletions specific to mammals

Small deletions at sequences with MH are an additional feature of ID4 (Fig. 1a) that is not observed in rnh201Δ plo2-M644G yeast (Fig. 1b). However, consistent with a ribonucleotide-induced mutational origin in mammalian cells, MH deletions are observed frequently in RNase-H2-deficient RPE-1 cells, in which SNMH sites account for 31% of 2 bp deletions, indicating that, in humans, deletions at SNMH sites share the same aetiology as deletions occurring at SSTRs. Taken together, our reporter and mutation-accumulation experiments demonstrate that genome-embedded ribonucleotides cause a similar mutational signature in yeast and mammalian cells. Thus, topoisomerase-1-mediated mutagenesis probably also occurs in humans and is associated with 2–5 bp deletions at SSTR and SNMH sequences.

ID4 mutations in a mouse cancer model

To determine whether TOP1-induced mutations resulting in the ID4 signature can be detected in vivo, we next studied a RER-deficient mouse cancer model in which Villin-cre conditional deletion of Rnaseh2b and Tp53 results in intestinal malignancy57. WGS analysis of paired tumour–normal tissue samples from 6 mice identified a total of 989 high-confidence tumour-specific somatic indels. An analysis of the resulting mutational signature established that ID4 substantially contributed in all tumours (Fig. 4a, b and Extended Data Fig. 6a), accounting for 32% of acquired indels. Consistent with a transcription-associated process, the ID4 signature was again most evident in transcribed genomic regions (Fig. 4b). The commonly occurring cancer signatures1 ID1, ID2 and ID5 were also observed, consistent with expectations of multiple mutational processes active in neoplasia.

The observed ID4 mutation spectrum corresponded closely to that observed in the RPE-1 mutation-accumulation experiment: 28% of indels were at 2–5 bp deletions, of which the majority were again 2 bp deletions (82%) predominantly at SSTRs (51%) and sites of SNMH (34%) (Extended Data Fig. 6b, c). This is consistent with the occurrence of TOPI-induced somatic mutations at genome-embedded ribonucleotides in vivo, conserved across different tissue and cellular contexts, and shows that this process can be detected in a cancer setting.

A sequence motif for ID4 mutations

Although COSMIC defines the ID4 signature on the basis of indel size and repeat/MH context (Fig. 1a), the number of indels in the mouse
Fig. 4 | RER-deficient tumours have an ID4 signature associated with transcription and a TNT sequence motif. a, ID4 contributes substantially to the mutational spectrum of Rnash2b-KO mouse intestinal tumours (WGS, paired tumour-normal samples from \( n = 6 \) mice). b, ID4 contribution is greater in transcribed regions of the genome. Statistical analysis was performed using a two-sided Fisher’s exact test, comparing ID4 versus other indels. \( n = 969 \) indels from 6 biologically independent tumours. c, 2 bp STR/SNMH deletions have biased sequence composition. Genome, frequency of dinucleotides in indels from 6 biologically independent tumours. d, Two-sided Fisher’s exact test, comparing ID4 versus other indels. e, A TNT sequence motif is present at all 2 bp STR and SNMH deletions. d, Sequence logo: two-bit representation of the sequence context of 2 bp deletions at STR and SNMH sequences. e, Deletion sites are significantly enriched for the TNT sequence motif compared with genome-wide occurrence, for all genome sequences, as well as STR and SNMH sites. Statistical analysis was performed using two-sided Fisher’s exact tests, comparing observed versus expected. \( n = 228 \) (all, \( P = 1.7 \times 10^{-20} \)), \( n = 124 \) (STR, \( P = 0.0008 \)), \( n = 77 \) (SNMH, \( P = 1.4 \times 10^{-7} \)) deletions in 6 biologically independent tumours. f, Model for TOP1-mediated mutations at TNT sequences containing embedded ribonucleotides, in which strand realignment results in a two-nucleotide deletion (see main text). nt, nucleotide.

RER-deficient tumour model enabled us to further investigate the characteristics of mammalian topoisomerase-1-induced mutations. We focussed our analysis on 2 bp deletions, as such events represented 81% of >1 bp deletions in the context of tandem repeats and 85% of deletions in sequences with MH.

First, we classified all 2 bp deletions at STR/SNMH sequences into six non-redundant dinucleotide classes, grouping together complementary sequences (Fig. 4c). We noted that the deleted sequences substantially deviated from genome-wide frequencies, with a complete absence of CC/GG and CG/GC deletions, as well as an overrepresentation of the CT category (containing CT, TC, GA and AG deletions). All of the observed deletions therefore included at least one thymidine (T), which functionally could be accounted for by the very strong preference of mammalian topoisomerase-1 to cleave at a phosphodiester bond with a T immediately upstream.

Next, to investigate the wider sequence context, we aligned sequences containing all 228 two-bp deletions (Extended Data Fig. 6d), which indicated that deletions preferentially occur when T nucleotides are spaced at a two-base interval. Indeed, this TNT motif was present in 100% of SNMH (\( n = 77 \)) and STR sites (\( n = 124 \)), providing a common unifying sequence context for both deletion types (Fig. 4d), a finding that was replicated in both our RPE-1 (Extended Data Fig. 6e) and yeast datasets (Extended Data Fig. 7). We found that TNT is substantially over-represented at deletion sites compared with the genome-wide null expectation. Furthermore, although the TNT motif is common at tandem repeat sequences, 2 bp deletions at this motif are still significantly enriched when considering the occurrence of 2 bp STR and SNMH sequences in mouse and human genomes (Fig. 4e and Extended Data Fig. 6f), and STR sequences in the yeast genome (Extended Data Fig. 7).

A model for TOP1-mediated deletions

To account for thymidines spaced at a two-base interval and the occurrence of mammalian SNMH deletions, we developed a revised model based on the established strand realignment model for yeast Top1-mediated mutagenesis. In this ‘TNT model’, TOP1 cleaves preferentially 3’ of an embedded ribouridine, with nucleophilic attack by the 2’-OH of the ribose ring resulting in TOP1 release and formation of a non-ligatable nick with a terminal 2’,3’-cyclic phosphate (Fig. 4f (i–iii)). This then provides a substrate for TOP1 cleavage 2 bp or more upstream, preferentially at a thymidine. When this second cleavage event happens at a base that is identical to that of the first cleaved nucleotide—an event that is more likely at STR and MH sequences—strand realignment can then occur, resulting in a nick that is permissive to religation and TOP1 cleavage complex (TOP1cc) reversal (Fig. 4f (iv–vi)). An alternative mechanism of sequential Top1 cleavage, in which
double-strand breaks occur due to nicking of opposite strands\textsuperscript{31} could not be reconciled with our TNT model, but may account for deletions occurring at non-STR/SNHM sites. Within the TNT motif, deletions were most common at CT and GT dinucleotides in both mammals and yeast (Fig. 4c and Extended Data Figs. 6 and 7b, e), which may be explained, at least in part, by preferential incorporation of ribouridine at CT and GT dinucleotides\textsuperscript{32} (Extended Data Fig. 7f).

Implicating TOP1-TAM as the cause of the ID4 signature enables us to include additional features in the definition of this COSMIC signature, namely preference for a TNT sequence motif at 2 bp deletion sites and enrichment in transcribed genes. Hereafter, we refer to this extended definition as ID-TOP1. To establish the relevance of the ID-TOP1 signature for human disease and genetic variation, we next examined publicly available datasets.
ID-TOP1 in human cancer

RNAseH2B is frequently deleted in human cancer, in particular, in chronic lymphocytic leukaemia (CLL), given its proximity to a tumour suppressor locus, the DLEU2-miR-15a/miR-16-1 microRNA cluster33. Such RNase-H2-deficient human cancers should therefore be enriched for the ID4/ID-TOP1 signature. We analysed WGS data for 348 patients with CLL from two independent cohorts4,35, stratified on the basis of RNAseH2B-deletion status. Somatic variant calling identified a significant increase in 2–5 bp deletions in RNAse-H2-null tumours (Fig. 5a), whereas other indels were equally represented across wild-type, heterozygous and null categories (Extended Data Fig. 8a). Of the 2–5 bp deletions in tumours with biallelic RNAseH2B loss, more than half (57%) were 2 bp deletions, which were predominantly at STR and SNMH sequences and substantially enriched for the TNT motif (Extended Data Fig. 8b), consistent with the ID-TOP1 mutational signature. Furthermore, mutational signature decomposition for RNAse-H2-null CLL cases confirmed the presence of the ID4 signature, most apparent in genomic regions (Extended Data Fig. 8d). We therefore conclude that the ID-TOP1 signature is present in human cancer and enriched in tumours that are RNaseH2-deficient.

Topoisomerase 1 also causes mutations in RER-proficient cells10,11 (Fig. 1d–f) and is therefore likely to cause mutations in other cancers, with deletions expected to occur most frequently in highly transcribed genes1. Accordingly, analysis of WGS data across cancer types (International Cancer Genome Consortium (ICGC)/Pan-Cancer Analysis of Whole Genomes (PCAWG)) demonstrated that the 2–5 bp deletion rate correlates with expression levels of ubiquitously expressed genes (Pearson’s r = 0.86, P = 0.0014), with deletions markedly elevated in the most highly expressed genes (Fig. 5b), consistent with previous reports of such deletions in certain cancer genes36–38. Examination of 2 bp deletions (42% of 2–5 bp deletions) across cancer types also demonstrated them to be predominantly in STR and SNMH contexts (Extended Data Fig. 5f) and enriched for the TNT sequence motif (Fig. 5c). Furthermore, using a dataset of TOP1 cleavage events captured by TOP1-seq38, we found that 2–5 bp deletions increase in frequency with TOP1 enzymatic activity, with such deletions more prevalent in regions of high TOP1 activity (Fig. 5d). Similarly, TOP1-ID deletion rates also corresponded to TOP1 activity and transcription level, in contrast to all other deletions (Extended Data Fig. 5g, h). Taken together, these establishes a substantial role for TOP1-mediated mutagenesis in the generation of somatic deletions.

To further examine the role of transcription in deletion mutagenesis of cancer genomes, we identified genes that are highly expressed, but only in certain tissues. For prostate adenocarcinoma, highly expressed prostate-restricted genes were significantly enriched for 2–5 bp deletion mutations compared with other genes in this cancer type, as well as the same genes in other cancers (two-tailed Fisher’s exact test, odds ratio = 3.5, P = 2.5 × 10⁻⁶ after Bonferroni correction; Extended Data Fig. 8i). Importantly, this analysis considers the same sets of genes between cancer types and therefore rules out sequence composition biases as a confounding effect for elevated ID-TOP1 mutagenesis in highly expressed genes. Extending this approach in an all-versus-all comparison between 8 cancer types and 17 tissues demonstrated specificity between high expression in a tissue of origin and enrichment for 2–5 bp deletions (Fig. 5e). These results extend the relevance of TOP1-mediated mutagenesis to other cancers, confirm the ID-TOP1 mutational signature to be transcription-associated and support the occurrence of TAM in humans.

TOP1-mediated deletions in the germline

TOP1 is ubiquitously expressed, so we reasoned that it could cause germline as well as somatic mutations. To investigate this possibility, we examined mutations from parent–child trio WGS studies in the Gene4Denovo database39. De novo mutations identified in such datasets represent germline events, as they occur in germ cells or during early embryonic cell divisions. Strikingly, 2–5 bp deletions were the largest category identified, accounting for 33% of the 40,936 de novo indels (Fig. 5f), and the majority of these were compatible with the ID-TOP1 signature. Analysis of 2 bp deletions (41% of 2–5 bp deletions) demonstrated that most occur at STR or MH sites (Extended Data Fig. 9a, b), with enrichment of the TNT sequence motif both genome-wide and in the context of STR/SNMH sites (Fig. 5g and Extended Data Fig. 9e). Similarly, for 3 and 4 bp deletions, respectively, TNNT and TNNNT motifs were significantly over-represented compared with genome-wide expectation (Extended Data Fig. 9d), supporting sequential TOP1 cleavage and strand realignment as the underlying cause. Consistent with TOP1-TAM aetiology, 2–5 bp deletion and ID-TOP1 deletion frequency correlated with transcript expression in male germ cells (Fig. 5h and Extended Data Fig. 9e). We therefore conclude that the ID-TOP1 mutational signature also occurs in the human germline, implicating TOP1-induced strand realignment mutagenesis as an important mutational process in mammalian cells.

Discussion

Here we establish a biological basis for the ID4 cancer signature5, experimentally demonstrating that it occurs in RNase-H2-deficient cells both in vitro and in vivo. This implicates TOP1-mediated cleavage at genome-embedded ribonucleotides as its cause. TOP1 is cell-essential in mammals and it is therefore not possible to similarly confirm a genetic dependency on TOP1 in human cells, as has been done in yeast42. However, conservation of this mechanism across eukaryotes is supported by us finding a topoisomerase 1-dependent TNT deletion motif that is present in both yeast and humans, and demonstrating that deletion frequency is dependent on human TOP1 activity levels. Previously published research also provides evidence for TOP1 mutagenesis at ribonucleotide sites in humans. The reversible transesterification reaction of type 1 topoisomerases is conserved from yeast to humans4, and human TOP1 has site-specific activity for ribonucleotides40, causing DNA breaks in mammalian RNAse-H2-deficient cells31. Furthermore, the generation of 2 bp deletions through sequential TOP1 cleavage at embedded ribonucleotides has been biochemically reconstituted with both human and yeast enzymes39,40.

We define additional features of this ID-TOP1 mutational signature, with deletions strongly enriched at TNT motifs in both yeast and mammals, a sequence context that is specific to topoisomerase 1 (Extended Data Fig. 7g, h) and deletions that are most frequent in highly transcribed regions. As a consequence, we show that a TAM process that was first identified in yeast42,43 is relevant to higher eukaryotes, establishing TOP1-induced mutagenesis as an important process for human variation and disease. Additional signatures associated with topoisomerases or indeed RNAse H2 may be identified in the future, particularly given that ID17 has recently been linked to TOP2A-17D33 cancers45.

The substantial contribution of ID-TOP1 deletions to germline mutagenesis has particular importance given that such deletions will be disproportionately disruptive, particularly in transcribed regions. Notably, such deletions occur in the context of normal RER function, consistent with the mutagenic potential of topoisomerase 1 in physiological wild-type settings38,44 (Fig. 1d). Given that genome-embedded ribonucleotides are the most common endogenous lesion in replicating mammalian cells39, they are the most likely sites of TOP1-TAM mutagenesis, where TOP1 could cleave before their removal by RNAse H2-dependent RER. Processing of TOP1cc may be an alternative, less frequent source of 2–5 bp deletions40, but we did not detect ID4 in cancers treated with topoisomerase I inhibitor (Extended Data Fig. 8j). The canonical function of TOP1 is to relieve DNA topological stress, arising during both transcription and replication46 (Extended Data Fig. 10). Thus, TOP1-mediated deletions are not restricted to transcribed regions
of the genome, with deletions also evident in non-genic regions with high TOP1 activity (Extended Data Fig. 8k). However, overall, enhanced TOP1 activity associated with transcription accounts for more frequent mutagenesis within genes.

Given the essential nature of topoisomerase activity across tissues and cell states, TOP1-mediated mutagenesis probably occurs in many contexts. The frequent TOP1-mediated human germline mutations (Fig. 5l–k) and the identification of ID4 at early embryonic stages suggest developmental vulnerability to TOP1-TAM. Moreover, 2–5 bp somatic deletions at SSRs are also observed at high frequency in non-dividing neurons, and ID4 has been identified in multiple tumour types. As such, this mutational process is likely to be important not only in cancers with RER deficiency, but also those with high TOP1 activity and tumours with defects in relevant repair mechanisms, such as enzymes that process TOP1 cleavage or non-ligatable TOP1-induced nicks. Furthermore, alternative RER pathways may exist that could reduce TOP1 mutagenesis. The ID-TOP1 signature may provide a useful biomarker with potential future diagnostic and therapeutic use, for example, as an indicator of TOP1-induced genomic instability targetable by PARP or ATR inhibitors.

In conclusion, alongside its essential role in relieving DNA torsional stress, TOP1 also drives mutagenesis in somatic and germline contexts, relevant to neoplasia, inherited disease and human variation.

Online content
Any methods, additional references, Nature Research 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/s41586-022-04403-y.

1. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat. Rev. Mol. Cell Biol. 9, 958–970 (2008).
2. Datta, A. & Jinks-Robertson, S. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science 266, 1616–1619 (1995).
3. Herman, R. K. & Dworkin, N. B. Effect of gene induction on the rate of mutagenesis by ICR-191 in Escherichia coli. J. Bacteriol. 106, 543–550 (1971).
4. Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. Annu. Rev. Genet. 48, 341–359 (2014).
5. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 101–120 (2020).
6. Pommer, Y., Sun, Y., Huang, S. N. & Nitis, J. L. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nat. Rev. Mol. Cell Biol. 17, 703–721 (2016).
7. Kunkel, T. A. Evolving views of DNA replication (m)diversity. Cold Spring Harb. Symp. Quant. Biol. 74, 91–101 (2009).
8. Ciccio, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. Mol. Cell 40, 179–204 (2010).
9. Lippert, M. J., Freedman, J. A., Barber, M. A. A. & Jinks-Robertson, S. Identification of a distinctive mutation spectrum associated with high levels of transcription in yeast. Mol. Cell. Biol. 24, 4801–4809 (2004).
10. Lippert, M. J. et al. Role for topoisomerase 1 in transcription-associated mutagenesis in yeast. Proc. Natl Acad. Sci. USA 108, 698–703 (2011).
11. Takahashi, T., Burguiere-Slezak, G., Auffret Van Der Kemp, P. & Boiteux, S. Topoisomerase I provokes the formation of short deletions in repeated sequences upon high transcription in Saccharomyces cerevisiae. Proc. Natl Acad. Sci. USA 108, 693–697 (2011).
12. Kim, N. et al. Mutagenic processing of ribonucleosides in DNA by yeast topoisomerase I. Science 332, 1561–1564 (2011).
13. Nick McElhenny, S. A. et al. Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. Proc. Natl Acad. Sci. USA 107, 4949–4954 (2010).
14. Reijns, M. M. et al. Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. Cell 149, 1008–1022 (2012).
15. Sparkes, J. L. et al. RNase H2-initiated ribonucleotide excision repair. Mol. Cell 47, 980–988 (2012).
16. Huang, S. N., Ghosh, S. & Pommier, Y. Topoisomerase I is alone sufficient to produce short DNA deletions and can also reverse nicks at ribonucleotide sites. J. Biol. Chem. 290, 14068–14076 (2015).
17. Sparkes, J. L. & Burgers, P. M. Error-free and mutagenic processing of topoisomerase I-provided damage at genomic ribonucleotides. EMBO J. 34, 1259–1269 (2015).
18. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 126–133 (2016).
19. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr. Opin. Genet. Dev. 24, 52–60 (2014).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
Georgia C. Chan14, Greg Elgar14, Tom Fowler14, Adam Giess14, Angela Hamblin14, Shirley Henderson14, Tim J. P. Hubbard14, Rob Jackson14, Louise J. Jones14, Dalia Kasperaviciute14, Melis Kayikci14, Athanasios Kousathanas14, Lea Lahnstein14, Sarah E. A. Leigh14, Ivonne U. S. Leong14, Javier F. Lopez14, Fiona Maleady-Crowe14, Meriel McEntagart14, Federico Minneci14, Loukas Moutsianas14, Michael Mueller14,15, Niurka Murugesa14, Anna C. Need14,15, Peter O’Donovan14, Chris A. O’dhams14, Christine Patch14, Mariana Buongermino Pereira14, Daniel Perez-Gil14, John Pullinger14, Tahrima Rahim14, Augusto Rendon14, Tim Rogers14, Kevin Savage14, Kushmita Sawant14, Richard H. Scott14, Afshan Siddiq14, Alexander Sieghart14, Samuel C. Smith14, Alona Sosinsky14,15, Alexander Stuckey14, Mélanie Tanguy14, Ana Lisa Taylor Tavares14, Ellen R. A. Thomas14,15, Simon R. Thompson14, Arianna Tucci14,15, Matthew J. Welland14, Eleanor Williams14, Katarzyna Witkowska14 & Suzanne M. Wood14,15

14Genomics England, London, UK. 15William Harvey Research Institute, Queen Mary University of London, London, UK.

Colorectal Cancer Domain UK 100,000 Genomes Project

Daniel Chubb9, Alex Cornish9, Ben Kinnersley9, Richard Houlston9, David Wedge14, Andreas Gruber16, Anna Frangou17, William Gross9, Trevor Graham9, Andrea Sottoriva14, Gulio Caravagna9, Nuria Lopez-Bigas20, Claudia Arnedo-Pac20, David Church17, Richard Culliford9, Steve Thom9, Phil Quirke21, Henry Wood9, Ian Tomlinson6 & Boris Noyvert14

9Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK. 10Wellcome Centre for Human Genetics, Oxford, UK. 11Cancer Institute, University College London, London, UK. 12Barts Cancer Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. 13Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. 14Pathology and Data Analytics, Leeds Institute of Medical Research, St James’s University Hospital, University of Leeds, Leeds, UK.
Methods

Plasmids
A description of all plasmids used in this work is provided in Supplementary Table 1. The *S. cerevisiae* reporter was generated by DNA synthesis (GeneArt Gene Synthesis, Thermo Fisher Scientific; gBlocks Gene Fragments, IDT) and conventional cloning (restriction, ligation and Quikchange site-directed mutagenesis). The final construct (pTCW12) was used for *S. cerevisiae* reporter strain construction and fluctuation assays. A Gateway compatible reporter construct for mammalian cells (pTCW14) was generated in a similar manner using a combination of DNA synthesis and conventional cloning strategies. Gateway cloning was then used to move the reporter cassette into pAAVS-Nst-CAG-Dest (a gift from K. Woltjen; Addgene plasmid, 90439) to generate pTCW15 for targeting it to the human AAVS1 locus.

In silico redesign of the hygromycin-resistance gene
To increase the frequency of 2 bp tandem repeats, synonymous substitutions were introduced into the 1 kb *bph* coding sequence, the Klebsiella pneumoniae hygromycin-resistance gene (HygR)52. Using Python, a 5-codon (15-base) sliding window was moved one codon at a time, to identify all possible synonymous permutations. Permutations were ranked on the basis of tandem dinucleotide repeat sequence length, with the highest ranking sequences used to replace whole codons, prioritising dinucleotide repeats over mononucleotide repeats. Edited codons were then censored from subsequent permutation. Subsequently, to eliminate stop codons that would arise after a 2 bp deletion or equivalent frameshift mutations, further synonymous changes were made, where possible preserving tandem repeat sequences.

Yeast strains and growth conditions
All *S. cerevisiae* strains used in this research (Supplementary Table 2) are isogenic by BY4741 (ref. 53) and were grown at 30 °C. TOP1 and RNH201 open reading frames were deleted using one-step allele replacement using PCR products generated from plasmid templates with selection cassettes (Supplementary Table 2) and primers containing 60-nucleotide homology directly upstream and downstream of the open reading frame. Gene deletions were confirmed by PCR. The 2 bp deletion reporter was inserted at the *AGP1* locus using a PCR product amplified from pTCW12 using primers AGP1-MX6-F and AGP1-MX6-R (Supplementary Table 3). The correct reporter insertion was confirmed by PCR and Sanger sequencing. Growth under selection was on YPD medium (10 g 1⁻¹ yeast extract, 20 g 1⁻¹ bactopeptone, 20 g 1⁻¹ dextrose, 20 g 1⁻¹ agar) supplemented with hygromycin B (300 mg l⁻¹), nourseothricin (100 mg l⁻¹) and/or G418 (1 g 1⁻¹), or on synthetic defined medium supplemented with hygromycin B (300 mg l⁻¹), nourseothricin (100 mg l⁻¹) and/or G418 (1 g 1⁻¹). A 1.3 kb region including HygroR was amplified in two overlapping amplicons (primers S297F and S113R; S752 and S1658R) using FastStart PCR MasterMix (Roche) and direct colony PCR (5 min at 95 °C; then 35 cycles of 30 s at 95 °C, 30 s at 58 °C and 45 s at 72 °C; and then 45 s at 72 °C). Each amplicon was Sanger sequenced using the primers described in Supplementary Table 3, and analysed using Sequencer v.5.4.6 (Gene Codes Corporation) and/or Mutation Surveyor v.3.30 (SoftGenetics). Mutation rates (per bp) were calculated for 1,032 bp of sequence in which productive frameshift mutations can occur.

Cell lines
A summary of the human cell lines used in this research is provided in Supplementary Table 4. All cells were grown at 37 °C and 5% CO₂, authenticated using STR DNA profiling in the laboratories of origin and shown to be mycoplasma negative through routine testing. HeLa cells (a gift from G. Stewart, University of Birmingham, UK; originally purchased from ATCC) were grown in Dulbecco’s Modified Eagle Medium (DMEM; Gibco/Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS), 50 U ml⁻¹ penicillin and 50 μg ml⁻¹ streptomycin. hTERT RPE-1 cells (a gift from D. Durocher, University of Toronto, Canada; originally purchased from ATCC) were grown in DMEM/F12 medium mixture (Gibco/Thermo Fisher Scientific) supplemented with 10% FBS, 50 μl ml⁻¹ penicillin and 50 μg ml⁻¹ streptomycin. The 2 bp deletion reporter was integrated at the AAVS1 safe harbour locus in HeLa cells using a published CRISPR–Cas9 targeting protocol51. HeLa cells were transfected with pXAT2 and pTCW15 in Opti-MEM reduced-serum medium using Invitrogen Lipofectamine 2000 (Thermo Fisher Scientific). After 48 h cells were replated in medium containing 500 μg ml⁻¹ G418 and, after another 48 h and a second round of replating in selective medium, single cells were sorted into 96-well plates using a BD FACSAjax instrument (BD Biosciences). The resulting G418-resistant clones were screened by PCR for reporter integration at the correct locus, retention of integration-free AAVS1 and Sanger sequencing of resulting PCR products. Single-locus integration was confirmed by fluorescence in situ hybridization (FISH) as previously described52, using pTCW16 to generate a fluoroscently labelled probe. The full reporter sequence of selected clones was checked, with amplification of a 1.9 kb fragment using Prime Star Max PCR Master Mix (Takara Bio) with the primers HygroR_up and Puror_rev (40 cycles of 10 s at 98 °C, 15 s at 70 °C and 2 min at 72 °C), followed by Sanger sequencing with additional primers (Supplementary Table 3). To generate RNASEH2A-KO reporter cells, the selected parental HeLa reporter clone was transfected with pMAR26 and pMAR27 (Supplementary Table 1), using Lipofectamine 2000. Then, 48 h after transfection, single EGFP-expressing cells were sorted into 96-well plates and grown until colonies formed. Initial screening was based on PCR amplification (primers RNASEH2A-ex1F and RNASEH2A-ex1R) of the CRISPR–Cas9-targeted region of RNASEH2A with mutations present in selected clones determined by Sanger sequencing. The cellular RNase H2 status was then confirmed by immunoblotting, RNase H2 enzymatic activity assay and alkaline gel electrophoresis to determine ribonucleotide content of genomic DNA (detailed methods are provided below).

Fluctuation assays in human
Hygromycin-resistant HeLa reporter cells (400 μg ml⁻¹ hygromycin B) were recovered from frozen stocks in the absence of selection. The next day, 10 wells of a 96-well plate were seeded with 2,000 cells per well for each line. The experiment was performed with the operator blinded to the identity of the cell lines. Cells were cultured under non-selective conditions and replated subsequently in 24-well, 6-well plates and ultimately 175 flasks, in which they were grown to confluence. Cells were then dissociated using Gibco TrypLE (Thermo Fisher Scientific)

Fluctuation assays in yeast
Fluctuation assays were performed as previously described44. Yeast was grown overnight in YPD medium with hygromycin B (300 mg l⁻¹), plated on YPD and grown at 30 °C to obtain individual colonies derived from a single cell without HygroR mutations. For each strain, 16 independent colonies were then used to inoculate 5 ml YPD, and grown for 3 days at 30 °C with shaking at 250 rpm. Cells were pelleted by centrifugation and resuspended in 1 ml of H₂O. Undiluted suspensions for each culture were plated (100 μl per plate) on 2 YPD plates supplemented with 1 μl 1⁻¹ G418, with the exception of nrh201Δ for which a 1 4 dilution was used. Furthermore, each suspension was serially diluted to 10⁻⁶ of which 100 μl per plate was spread on 2 YPD plates to estimate the total number of viable cells per culture. Plates were incubated at 30 °C for 2–3 days, and colonies counted. Mutation rates were determined in Microsoft Excel 2016 for each individual culture, and an overall rate for each strain calculated using the Lea Coulson method of the median45. The number of mutants for each culture was ranked, and those ranked 4th and 13th were used to calculate the rates that define the lower and upper limits of the 95% CI. A single G418-resistant colony for each independent culture was used to determine the spectrum of frame-shift mutations. A 1.3 kb region including HygroR was amplified in two overlapping amplicons (primers S297F and S113R; S752 and S1658R) using FastStart PCR MasterMix (Roche) and direct colony PCR (5 min at 95 °C; then 35 cycles of 30 s at 95 °C, 30 s at 58 °C and 45 s at 72 °C; and then 45 s at 72 °C). Each amplicon was Sanger sequenced using the primers described in Supplementary Table 3, and analysed using Sequencer v.5.4.6 (Gene Codes Corporation) and/or Mutation Surveyor v.3.30 (SoftGenetics). Mutation rates (per bp) were calculated for 1,032 bp of sequence in which productive frameshift mutations can occur.
and the cells were counted using a Moxi Z automated cell counter. After serial dilution, 1,000 cells were plated into two 10 cm plates for each culture and grown for 14 days to determine plating efficiency. All other cells were plated into two 10 cm plates, and 0.5 μg ml⁻¹ puromycin was added after 4 h, with the medium subsequently changed every 2–3 days for 14 days to remove dead cells and maintain a puromycin concentration of 0.5 μg ml⁻¹.

To establish mutation spectra, colonies were removed by scraping and then cultured in a 96-well plate. When confluent, cells were lysed with 75 μl DirectPCR Lysis Reagent (Viagen Biotech) and 0.4 mg ml⁻¹ PCR-grade Proteinase K (Roche), heating overnight at 55 °C followed by 45 min at 85 °C. Only one sample per independent culture was used for PCR amplification and Sanger sequencing to determine the nature of mutations in the HygroR coding sequence. A 1.24 kb region including HygroR was amplified with Prime Star Max PCR Master Mix (Takara Bio), HygroR_up and H1327R primers (40 cycles of 10 s at 98 °C, 15 s at 70 °C, 2 min at 72 °C). Sanger sequencing was then performed with additional primers (Supplementary Table 3) and mutations identified using Mutation Surveyor v.3.3.0 (SoftGenetics). All mutants showed double traces of equal height from the point of indel mutations, consistent with the presence of two copies of the reporter in all reporter lines. As FISH indicated the presence of the reporter at a single A4VS1 locus, we inferred that two copies of the reporter were inserted in tandem at this locus. As a 2 bp deletion or equivalent frameshift mutation in either HygroR copy would bring the associated PuroR coding sequence into the translated reading frame, we corrected mutation rate calculations (per bp) for the presence of two copies.

To determine colony numbers, plates were washed with PBS, fixed with 2% formaldehyde in PBS for 10 min, rinsed with water and colonies were stained with 0.1% crystal violet solution for 10 min. Plates were then washed with water and left to dry before counting colonies. After counting, the experiment was performed in an unblinded manner. Mutation rates were determined for each individual culture in Microsoft Excel 2016, and an overall rate for wild-type and KO strains was calculated using the Lea Coulson method of the median. The number of mutants for each culture was ranked, and appropriate ranks were used to calculate the rates that define the lower and upper limits of the 95% CI.

Immunoblotting

Whole-cell extracts (WCE) to determine protein levels of RNase H2 subunits by immunoblotting and for RNase H2 activity assays were prepared as previously described76. Equal amounts of protein from WCE were separated by SDS–PAGE on 4–12% NuPAGE gels and transferred to PVDF. Membranes were probed in 5% milk (w/v; Marvel Original Dried Skimmed), TBS + 0.2% Tween-20 (v/v) with the following antibodies: sheep anti-RNase H2 (raised against human recombinant RNase H2, 1:10,000); mouse anti-RNase H2A G-10 (Santa Cruz Biotechnologies, sc-515475, A1416, 1:100); rabbit anti-GAPDH (Abcam, ab9485, 1:2,000, GR3380498-1). For detection, we used rabbit anti-sheep immunoglobulins/HRP (Dako, P04163, 0047199, 1:2,000; goat anti-mouse immunoglobulins/HRP (Dako, P0447, 20039214, 1:10,000); anti-rabbit IgG, HRP-linked antibodies (Cell Signaling Technologies, 7074S, 29, 1:10,000); Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences) and an ImageQuant LAS4000 device, or IRDye secondary antibodies and an Odyssey CLX Imaging System (LI-COR Biosciences). The uncropped immunoblots are presented in Supplementary Fig. 1.

RNase H2 activity assays

To assess cellular RNase H2 activity, a FRET-based fluorescence substrate release assay was performed as previously described44. In brief, RNase-H2-specific activity was determined by measuring the cleavage of double-stranded DNA substrate containing a single embedded ribonucleotide. Activity against a DNA-only substrate of the same sequence was used to correct for background activity. Substrates were formed by annealing a 3'-fluorescein-labelled oligonucleotide (GATCTGAGCC TGGGAAGC or GATCTGAGCTGGGAGCT; uppercase DNA, lowercase RNA) to a complementary 5'-DABCYL-labelled DNA oligonucleotide (Eurogentec). The reactions were performed in 100 μl reaction buffer (60 mM KCl, 50 mM Tris–HCl pH 8.0, 10 mM MgCl₂, 0.01% BSA, 0.01% Triton X-100) with 250 nM substrate in black 96-well flat-bottomed plates (Costar) at 24 °C. WCE was prepared as described above, protein concentrations determined using the Bio-Rad Bradford Protein Assay and the final protein concentration per reaction was 50 ng μl⁻¹. Fluorescence was read (100 ms) every 5 min for up to 90 min using the VICTOR2 1420 multilabel counter (Perkin Elmer), with a 480 nm excitation filter and a 535 nm emission filter. Initial substrate conversion after background subtraction was used to calculate RNase H2 enzyme activity.

Alkaline gel electrophoresis

To determine the presence of excess genome-embedded ribonucleotides in nuclear DNA, alkaline gel electrophoresis of RNase-H2-treated genomic DNA was performed as previously described18. In brief, total nucleic acids were isolated from pellets from around 1 million cells by incubation in ice-cold buffer (20 mM Tris–HCl pH 7.5, 75 mM NaCl, 50 mM EDTA) with 200 μg ml⁻¹ proteinase K (Roche) for 10 min on ice, followed by addition of Na-laurylsarcosine sodium salt (Sigma-Aldrich) to a final concentration of 1%. Nucleic acids were extracted using phenol–chloroform, then isopropanol-purified and dissolved in nucleic acid-free water. For alkaline gel electrophoresis, 500 ng of total nucleic acids was incubated with 1 pmol of purified recombinant human RNase H2 (isolated as previously described17) and 0.25 μg of DNase-free RNase (Roche) for 30 min at 37 °C in 100 μl reaction buffer (60 mM KCl, 50 mM Tris–HCl pH 8.0, 10 mM MgCl₂, 0.01% Triton X-100). Nucleic acids were ethanol-precipitated, dissolved in nucleic acid-free water and 250 ng was separated on 0.7% agarose gels in 50 mM NaOH, 1 mM EDTA. After overnight electrophoresis, the gel was neutralised in 0.7 M Tris–HCl pH 8.0, 1.5 M NaCl and stained with SYBR Gold (Invitrogen). Imaging was performed on a FLA-5100 imaging system (Fujifilm), and densitometry plots were generated using AIDA Image Analyzer v.3.44.035 (Raytest).

Mutation-accumulation experiment

TPS3-KO hTERT RPE-1 cells without and with loss-of-function mutations in RNASEH2A or RNASEH2B, introduced by CRISPR-Cas9 genome editing, a gift from D. Durocher, have previously been described33. RNase-H2-proficient (wild type), RNASEH2A-KO and RNASEH2B-KO cells were single-cell sorted into 96-well plates using the BD FACsJazz instrument (BD Biosciences). Multiple individual clones for each were expanded to confluent T75 flasks for cryopreservation and genomic DNA isolation of these ancestral populations. Moreover, lines were again single-cell sorted into 96-well plates to start the mutation-accumulation experiment. Cultures were expanded by subsequent growth in 24-well, 6-well plates and T75 flasks until confluent (approximately 25 population doublings), and this process of single-cell sorting and expansion was repeated four more times, providing bottlenecks to capture mutations that occurred since the previous sort. From the first to the last single-cell sort, a total of approximately 100 population doublings occurred and the final culture was expanded for cryopreservation and genomic DNA isolation of these end-point populations.

Genomic DNA was isolated using phenol extraction as previously described50, for alkaline gel electrophoresis and WGS. Library preparations and sequencing were performed by Edinburgh Genomics. Libraries were prepared using Illumina SeqLab specific TruSeq PCRFree High Throughput library preparation kits according to the manufacturer’s instructions, with DNA samples sheared to a mean insert size of 450 bp. Libraries were sequenced using paired-end reads on the Illumina HiSeqX instrument using v2.5 chemistry to achieve minimum mean genome-wide sequencing depth of 30× per sample.
Mouse WGS analysis

Villin-cre Trp53<sup>Δ/Δ</sup>Rnaseh2b<sup>Δ/Δ</sup> mice with epithelial-specific deletion of Trp53 and Rnaseh2b on a C57Bl/6 background have been described previously<sup>29</sup>. Animal experiments were conducted with appropriate permission, in accordance with guidelines for animal care of the Christian-Albrechts-University, in agreement with national and international laws and policies. No randomization or blinding was performed. Paired tumour–normal DNA was isolated from small intestinal tumours (Trp53<sup>Δ/Δ</sup> Rnaseh2b<sup>Δ/Δ</sup> ) and liver tissue (Trp53<sup>Δ/Δ</sup> Rnaseh2b<sup>Δ/Δ</sup> ) from female mice (aged 52 weeks), using the Qiagen DNeasy Blood & Tissue Kit. Library preparations and sequencing were performed by Edinburgh Genomics using Illumina DNA PCR-Free Library Prep according to the manufacturer's instructions. Paired-end sequencing was performed by Edinburgh Genomics on a NovaSeq 6000 using v1.5 chemistry. Mean genome-wide sequencing depth of at least 30× for liver samples and 60× for tumour samples was obtained.

5. cerevisiae WGS analysis

WGS SRA files for <i>rnh201Δ pol2-M644G</i> S. cerevisiae<sup>25</sup> from the NCBI Sequence Read Archive (SRA) were converted to FASTQ files using SRA Toolkit v.2.5.4-1 (SRA Toolkit Development Team; http://ncbi.github.io/sra-tools/). FASTQ reads were aligned to the SGE6939 1.03 ref_v2 reference genome<sup>49</sup> (Supplementary Table S5) and sorted BAM files were created using BWA-MEM (v.0.7.12)<sup>61</sup>, and deduplicated with SAMBLASTER (v.0.1.22)<sup>64</sup>. To select high-quality indel variants, GATK (v.3.6-0) Haplotype Caller (without base quality score recalibration)<sup>60</sup> variant calling was performed with ‘Hard Filters’ (‘filterExpression “QD < 2.0 || FS > 200.0 || ReadPosRankSum < −20.0”’). Filtering for strain-specific variants was performed as previously described<sup>60</sup>, with minor modifications. The filters were as follows: (1) eliminate variants shared with an ancestral clone; (2) required ≥20 reads for variant allele in descend; (3) exclusion of repetitive sequences as defined in ref. 60; and (4) reference/variant depth ratio 0.4–0.6, ≤0.4 if homozygous variant allele.

RPE-1 WGS analysis

FASTQ files were converted to unaligned BAM format and Illumina adapters were marked using GATK (v.4.1.9.0) FastqToSam and MarkIIlluminaAdapters tools<sup>62</sup>. Reads were aligned to the human genome (hg38, including alt, decoy and HLA sequences) using BWA-MEM (v.0.7.16)<sup>61</sup> and read metadata were merged using GATK’s MergeBamAlignment tool. PCR and optical duplicate marking and base quality score recalibration were performed using GATK. Variants from NCBI dbSNP build 151 were used as known sites for base quality score recalibration. Post-processed alignments were genotyped using Mutect2, Strelka2, Platypus and SvABA using somatic calling models for each pair of ancestral and end-point cultures, as described below.

Mouse WGS analysis

FASTQ processing and alignment were performed as for RPE-1 WGS analysis, using the GRCm38 mouse genome reference and known variant sites from the Mouse Genomes Project<sup>41</sup> (REL-1807-SNP_s, Indels) for base quality score recalibration. Somatic variant calling of post-processed alignments was performed using Mutect2, Strelka2, Platypus and SvABA for each tumour–liver pair, as described below. Samairal v.0.2.12 (https://github.com/brentp/somairal) was used to confirm each paired tumour and liver sample originated from the same animal.

Human ethics approval

Data generated from Genomics England 100,000 genomes and ICGC-CLL studies were analysed. In these respective studies, informed consent for participation was obtained. Ethical approval for Genomics England 100,000 genomes project: East of England and South Cambridge Research Ethics Committee; CLL-ICGC: International Cancer Genome Consortium (ICGC) guidelines from the ICGC Ethics and Policy committee were followed and the study was approved by the Research Ethics Committee of the Hospital Clinic of Barcelona.

CLL WGS analysis

Genomics England. CLL tumour–normal pairs (n = 198) were processed as part of the 100,000 Genomes Project (pilot and main programme v8). Samples were sequenced using the Illumina HiSeq X System with 150 bp paired-end reads at a minimum of 75× coverage for tumours and 30× coverage for germline samples. Reads were mapped to GRCh38 using IASAC aligner (v.03.16.02.19)<sup>69</sup>. Single-nucleotide variants (SNVs) and indels were called using Strelka v.2.4.7 using somatic calling mode. Structural and copy number variants were called using Manta (v.0.28.0) and Canvars (v.1.3.1)<sup>70</sup>, respectively. Samples with a tumour purity estimate from Canvas of less than 50% were excluded from analysis. RNASHEHZB copy number was determined using a combination of Canvas, Manta, read depth counts with samtools (v.1.9) and confirmed by manual inspection using IGV (v.2.5.0)<sup>68</sup>.

ICGC. WGS from the ICGC-CLL cohort (n = 150) was reanalysed. Raw reads were mapped to the human reference genome (GRCh37) using BWA-MEM (v.0.7.15)<sup>42</sup>. BAM files were generated, sorted and indexed, and optical or PCR duplicates were flagged using biobambambam2 (https://gitlab.com/german.tischler/biobambambam2, v.2.0.65). Copy-number alterations were called from WGS data using Battenberg (cppBattenberg, v.3.2.2)<sup>71</sup>, ASCAT (ascatNgs, v.4.1.0)<sup>70</sup>, and Genome-wide Human SNP Array 6.0 (Thermo Fisher Scientific) data<sup>43</sup> reanalysed using Nexus 9.0 Biodiscvery software (Biodiscovery). RNASHEHZB copy number was established by combining the three analyses and manual review with IGV.

Colorectal cancer WGS analysis

Irinotecan-treated (n = 39) and irinotecan-untreated (n = 78) colorectal cancers from the 100,000 Genomes Project Colorectal Cancer Domain were I×2 matched using a multivariate greedy matching algorithm without replacement, implemented in the Matching R-package<sup>72</sup>. Matching was conducted considering sex, age at sampling, whether a primary tumour or metastasis had been sequenced, microsatellite instability status, and whether the individual had previously received radiotherapy, oxaliplatin, capcetabine or fluorouracil treatment.

Somatic variant calling

Somatic variant calling was performed in parallel using four distinct methods: Mutect2 (as part of GATK v.4.1.9.0)<sup>73</sup>; Strelka2 (v.2.1.9.10)<sup>74</sup>; SvABA (v.1.1.3)<sup>35</sup> and Platypus (v.8.1.5)<sup>75</sup>. High-confidence indel calls were defined as the intersected output of these four tools, where variants passed all filters for ≥3 of 4 callers. The intersection was performed using the bcftools (v.1.10.2)<sup>76</sup> isec function after normalizing variant calls and left-aligning ambiguous alignment gaps using the bcftools norm function. For Platypus (v.8.1.5)<sup>75</sup> joint calling all samples in each cohort was performed before filtering for somatic variants; the other variant callers were run in paired tumour–normal mode. For the RPE-1 mutation-accumulation experiment the end-point and ancestral cultures were defined as ‘tumour’ and ‘normal’ samples, respectively. Variant filtering strategies were optimized to both available information on segregating genetic variation for humans and mice, and the functionality of each calling method as detailed below.

Mutect2. Unfiltered genotypes for all normal samples were combined to filter germline variants. Somatic calls were obtained using GATK’s FilterMutectCalls command. Human polymorphism data and allele frequencies from gnomAD<sup>46</sup> were provided to Mutect2 for the filtering of germline variants.

SvABA. Germline indel and structural variants were filtered using the --dbsnp-vcf and --germline-sv-database options. Mouse indels
were obtained from Mouse Genomes Project version 5 SNP (ftp://ftp-mouse.sanger.ac.uk/REL-1505-SNPs_Indels/mgp.v5.merged.indels.dbSNP142.normed.vcf.gz); structural variants from SV release version 5 (ftp://ftp-mouse.sanger.ac.uk/REL-1606-SV/mgpv5.SV_insertions.bed.gz and ftp://ftp-mouse.sanger.ac.uk/REL-1606-SV/mgpv5.SV_deletions.bed.gz). Human indels were extracted from NCBI dbSNP build 151 and common structural variants from dbVAR (https://hgdownload.soe.ucsc.edu/gbdb/hg38/bbi/dbVar/).

**Strelka2.** Candidate small indels for each pair were first generated by Manta (v1.6.0) in somatic calling mode. Strelka2 was then executed in somatic calling mode for each pair with Manta’s candidate small indels output provided to the –indelCandidates option.

**Platypus.** Germline variants were filtered on the basis of any normal sample with ≥2 variant allele reads. Somatic variant calls for each sample pair were retained if tumour:end-point sample > 2 variant reads; site depth > 9; and normal sample read depth ≥ 20, < 2 variant reads. Moreover, a >10× ratio of tumour to normal for variant/total depth was required.

For Genomics England CLL tumour–normal pairs, pre-existing Strelka2 calls from the 100,000 Genomes Project pipeline were used, while variant calling with Mutect2, Platypus and SvABA was performed as above. Colorectal cancer tumour–normal pairs from Genomics England were processed as for Genomics England CLL but without Mutect2 analysis. For ICGC CLL, somatic indels were called using Mutect2 (v.4.0.2.0)\(^2\), Strelka2 (v.2.8.2)\(^3\), SvABA (v.1.10)\(^4\) and Platypus (v.0.8.3)\(^5\). Candidate small indels generated by Manta (v1.2)\(^6\) were used as input for Strelka2. Mutect2, Strelka2 and SvABA were run in paired tumour–normal mode. somaticMutationDetector.py (https://github.com/andyrimmer/Platypus/tree/master/extensions/Cancer) was used to identify somatic indels called by Platypus with a minimum posterior of 1. SNVs called by Platypus were considered to be somatic if they had at least 2 alternative reads in the tumour, fewer than 2 alternative reads in the normal, a minimum tumour VAF of 10× the control VAF, and a minimum depth of 10.

**Germline mutation analysis**

De novo WGS variants were downloaded from the Gene4Denovo database (Supplementary Table 5). Reference assembly conversion errors were removed by discarding variants for which the reference allele did not match the genome reference at the given position or for which the variant position was greater than the length of the reference chromosome. Furthermore, individuals with total de novo variants below the 10th (n = 33) or above the 90th (n = 140) percentile were excluded. For germline gene expression we used predefined expression groups\(^6\) based on Ensembl release 90 annotation (ftp://ftp.ensembl.org/pub/release-90/gtf/homo_sapiens/Homo_sapiens.GRCh38.90.gtf.gz). Initially stratified as nine expression groups from 1 (unexpressed) to 9 (high), we collapsed them into a smaller set of unexpressed (1), low (2, 3, 4), mid (5, 6, 7) and high (8, 9). The annotations were converted to GRCh37 coordinates using liftover (kent source v.417). Genomic segments overlapping multiple distinct expression groups, due to overlapping genes, were assigned to the higher of those expression groups. For each expression group, we summed the count (c) of de novo indels contained within the genomic span of those genes. This was converted to rate estimates by dividing by the union genomic span (g nucleotides) of genes in that expression group, and adjusting for the number of mutated genomes considered (n); rate = c / (g*n). To obtain 95% CIs, gene selection was bootstrapped (sampled to an identical number with replacement) 100 times and the 0.025 and 0.975 quantiles of the bootstrapped rate calculation taken as the 95% CI.

**ICGC pan-cancer expression analysis**

The ICGC PCAWG somatic mutations\(^6\) were downloaded from the NCBI Gene Expression Omnibus database (accession code GSE57628, samples GSM31385717 and GSM31385718). Autosomal chromosomes 1 to 22 and the X chromosome were divided into 1 kb bins and, for each bin, the amount of mappable sequence was determined using Umap’s regions mappable using 36-mers\(^8\) to approximate the read length of the TOP1-seq data. For each 1 kb window, the TOP1-seq signal within mappable regions was summed for each replicate and the mean signal was calculated. This mean was divided by the amount of mappable sequence to calculate the TOP1-seq signal per bp and each 1 kb window was then assigned to decile bins using this value.

**ICGC pan-cancer TOP1-seq analysis**

Data corresponding to two replicates of TOP1-seq, a modified ChIP-seq technique to immunoprecipitate only catalytically engaged TOP1 (ref. \(^9\)), were downloaded from the NCBI Gene Expression Omnibus database (accession code GSE57628, samples GSM31385717 and GSM31385718). Autosomal chromosomes 1 to 22 and the X chromosome were divided into 1 kb bins and, for each bin, the amount of mappable sequence was determined using Umap’s regions mappable using 36-mers\(^8\) to approximate the read length of the TOP1-seq data. For each 1 kb window, the TOP1-seq signal within mappable regions was summed for each replicate and the mean signal was calculated. This mean was divided by the amount of mappable sequence to calculate the TOP1-seq signal per bp and each 1 kb window was then assigned to decile bins using this value.

Somatic deletion calls from ID4-positive PCAWG samples (as defined in https://dcc.icgc.org/api/v1/download?fn=/PCAWG/mutational_signatures/Signatures_in_Samples/SP_Signatures_in_Samples/...
PCAWG_SigProfiler_ID_signatures_in_samples.csv) were counted within the same 36-mer mappable regions for each 1 kb window and either stratified by type (1 bp deletion, 2–5 bp deletion) or by the TN\textsuperscript{T} motif defined below. Relative rates of deletions in each category were calculated relative to the first TOP1-seq signal decile.

Mutational signatures
De novo extraction and decomposition of mutational signatures was performed in Python v.3.8.5 using SigProfilerExtractor (v.1.1.0)\textsuperscript{5}, along with SigProfilerMatrixGenerator (v.1.1.14/1.1.15)\textsuperscript{83} and SigProfilerPlotting (v.1.1.27). The recommended default settings (including 500 NMF replicates) were applied (https://github.com/AlexandrovLab/SigProfilerExtractor). Subtraction of mutations in RPE-1 wild-type cells from those detected in RNase H2 null cells was performed as follows. The average number of indels per line for each of the S3 categories was determined for the three wild-type lines. Counts per category for AKO and BKO lines were subtracted using these averages, with negative values set to 0. SigProfilerExtractor was then performed on the resulting WT-subtracted AKO and BKO ID-83 matrices for both de novo signature detection and decomposition analysis.

Indel sequence context analysis
WGS indels were categorized on the basis of repeat sequence context. Genome-wide occurrence of short repeats and regions of MH were identified and filtered to include only the mappable genome, defined by Umap’s regions mappable using 100-mers\textsuperscript{36}. For both WGS-identified indel variants and genome-wide occurrence, scoring of 2 bp deletions compliant with the TNT motif at MH/SSTR sites required the deleted bases to match the sequence NT with a T immediately 5′ of the deleted dinucleotides. More generally, for varying sized deletions these were considered to fit a TN\textsuperscript{T} motif if the deletion lay within an SSTR or region of MH containing the motif TN\textsubscript{r–1}T, where \(d\) is the length of the deletion. Genome-wide occurrences were estimated from 100,000 randomly generated deletions of given lengths within the mappable genome. For SSTRs and MH regions, all regions containing the respective motifs (TN\textsubscript{r–1}T or TN\textsubscript{r–1}A) were identified (where \(r\) is the length of the repeat unit and \(n > 1\)), and the fraction of SSTR/MH sequence containing TNT motifs was determined against total SSTR/MH sequence in the mappable genome.

To derive a null expectation for de novo deletions matching the TNT, TNNT and TNNNT motif for 2, 3 and 4 bp deletions, respectively, deletions at repeats from the Gene4Denovo database were first classified by deletion length, repeat type (STR or MH) and repeat length. Bootstrap samples of corresponding repeats from the genome were generated with 1,000 replicates. That is, for each deletion category an equal number of repeats of matching repeat type, repeat unit length and total repeat length were randomly drawn from the genome for each bootstrap sample.

Sequence logos
Genomic sequences containing 2 bp deletions were reversed and complemented when the deleted dinucleotide contained an adenosine (A), except when the dinucleotide was AT or TA. For SNMH and STR deletions, the position of the deleted dinucleotide cannot be unequivocally assigned and, therefore, the deleted sequence was right-aligned in the repeat/MH region, either to the most 3′ T, where present, or otherwise to the limit of the repeat/MH region. Sequences were converted to bit score matrices and logos were drawn using Logomaker (v.0.8)\textsuperscript{84}.

Embedded ribonucleotide sequence context analysis
EmRiboSeq data from rmh201A yeast prepared during mid-log phase growth\textsuperscript{25} were obtained (Supplementary Table 5) and aligned to the sac-cer3 reference genome as previously described to identify the genomic coordinates of genome-embedded ribonucleotides\textsuperscript{86}. The Bedtools (v.2.30.0)\textsuperscript{87} utilities groupby, slop and getfasta were used to extract and count the sequence context of genome-embedded ribonucleotides with downstream analysis and plotting implemented in R (v.4.0.5).

Genome sequence composition-adjusted relative rates were calculated as previously described\textsuperscript{2} such that, under the null expectation of no sequence bias in ribonucleotide incorporation, all sequence contexts have an expected relative rate of \(1/n\) where \(n\) is the number of contexts considered.

Statistical methods
Statistical testing was performed using GraphPad Prism v.9.1.1, Python v.3.8.5 or R v.3.3.1. Two-sided non-parametric Mann–Whitney \(U\) tests were performed for quantitative measurements; multiple testing correction, FDR set at 0.05; and, for categorical data, Fisher’s exact tests were performed in Python using stats.fisher_exact from scipy v.1.6.3. Calculation of cosine similarities was performed as follows. Mutations for each strain were converted into a vector, with ordered values representing different mutation categories as a proportion of total mutations. These were then compared in a pairwise manner. Given two vectors \(\textbf{A}\) and \(\textbf{B}\), the cosine similarity \((\cos(\theta))\) was calculated as:

\[
\cos(\theta) = \frac{\sum_{i}^{n} \textbf{A}_i \textbf{B}_i}{\sqrt{\sum_{i}^{n} (\textbf{A}_i)^2} \sqrt{\sum_{i}^{n} (\textbf{B}_i)^2}}
\]

Hierarchical clustering used the hclust function of R (v.4.1.0) with complete linkage clustering of pairwise cosine distances (1 − cosine similarity) between ID-83 mutation spectra, with 41 categories of productive reporter framseshift mutations. For bootstrap support, \(n = 1,000\) bootstrap datasets were generated by sampling with replacement the mutations observed with a strain, for each strain, and then calculating the cosine distance and hierarchical clustering for each bootstrap dataset. Reported bootstrap scores are the percentage of bootstrap replicates showing hierarchical clustering of which supports the clustering to the right of the indicated position.

To test the significance of cosine similarities, we used a null model based on the Dirichlet-multinomial distribution. In brief, when comparing two mutation count vectors, with total mutations \(m_1\) and \(m_2\), over \(n\) mutation classes, we constructed a distribution of cosine values by comparing 10,000 simulated pairs of random vectors generated as follows. For each simulated pair, we sampled from a Dirichlet-multinomial distribution with the concentration parameters as a vector of ones of dimension \(n\), and number of trials as \(m_1\) for the first vector in the pair, and \(m_2\) for the second vector. The null distribution was obtained by computing the cosine similarity of the 10,000 pairs of mutation count vectors.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
RPE-1 mutation-accumulation experiment and mouse tumour WGS data are available from the European Nucleotide Archive under accession number PRJEB48753. All other data were previously published and the sources are cited in Supplementary Table 5.

Code availability
Code documented in the Methods is available online (https://git.ecdf.ed.ac.uk/ID-TOP1).

51. Oceguera-Vanex, F. et al. Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives. Methods 101, 43–55 (2016).
52. Grütz, L. & Davies, J. Plasmid-encoded hygromycin B resistance: the sequence of hygromycin B phosphotransferase gene and its expression in Escherichia coli and Saccharomyces cerevisiae. Gene 25, 179–188 (1983).
Extended Data Fig. 1 | ID4 is distinct from small deletion signatures of known aetiology. a, b, The mechanistic basis for many COSMIC indel signatures is unknown, with only 9 out of 18 having a proposed aetiology. ID2 (a) is attributed to DNA polymerase slippage⁸⁸,⁸⁹ and ID6 (b) to microhomology mediated end-joining (MMEJ) activity, associated with HR deficiency⁵,⁹⁰. c, d, Mechanism for these signatures supported by: impaired MMR promoting replication slippage mutagenesis in MLH1⁻/⁻ colonic organoids resulting in ID2 (and ID1) signatures (c); ID6 contributing substantially (along with ID8) to the indel signature in ovarian cancer, in which HR deficiency is common (d). Analysis of data from⁵ in c; data for 73 ovarian adenocarcinomas with ID6 contribution from ICGC⁵,⁶⁰ in d.
Extended Data Fig. 2 | Yeast and human frameshift mutation reporters detect indels at tandem repeats. a, Yeast reporter. Synonymous substitutions were made in the hygromycin resistance gene (HygroR), such that it contained many short 2 bp tandem repeats (SSTRs). Expression from the TEF promoter (PTEF) ensures a constitutive high level of transcription. Mutations within HygroR that result in a frameshift simultaneously put the HygroR coding sequence out of frame and the downstream neomycin resistance (NeoR) sequence in frame, allowing antibiotic selection of cells with such mutations. b, Top1-dependent 2 bp SSTR deletions occur in both WT and rnh201Δ (RNase H2 null) yeast, with the highest mutation rate for rnh201Δ (related to Fig. 1d). c–e, WT and rnh201Δ have similar spectra, and differ from top1Δ strains. Mutation spectra of neomycin resistant colonies. n, number of independent colonies sequenced. Other: complex indels, missense mutations or mutation not characterised (c). Tree for pairwise clustering with percent bootstrap support to the right of the indicated position, based on cosine scores calculated for mutation spectra (Fig. 1e) of the 41 mutation categories that give productive reporter frameshift mutations (d). Matrix of pairwise cosine similarities and P-values between reporter mutation spectra in different yeast strains. Darker blue indicates greater similarity; darker grey greater significance. Test statistic is the cosine similarity value for 41 mutation categories and the null hypothesis is that that the cosine value will be distributed according to the Dirichlet-multinomial model, as described in Methods. The test is one-sided and no adjustments were made for multiple comparisons (e), f, Null distribution for cosine pairwise vector comparisons for 41 and 83 mutation categories. Plots, cosine values for 10,000 randomly generated pairs of vectors of mutation spectra. Each vector contained 100 randomly assigned mutations (see Methods for further details). Cosine value thresholds indicated for P < 0.05 and P < 0.01. g, The human reporter is expressed from the ubiquitous CAG promoter (PCAG), and NeoR is replaced with the puromycin resistance gene (PuroR) to allow more rapid antibiotic selection in mammalian cell culture.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Validation and characterisation of RNASEH2A+ and KO HeLa reporter cells. a–c, Reporter integration at the AAVS1 locus and retention of a reporter-free locus with a 200 bp deletion at the target site was confirmed by PCR and Sanger sequencing. Green arrow head, specific PCR product. Representative of at least 2 independent experiments. d, e, FISH shows integration of the reporter (d) at a single AAVS1 locus (e). Representative image of approximately one hundred mitotic chromosome spreads in 3 independent experiments. SA, splice acceptor; T2A, self-cleaving peptide; pA, polyadenylation site; also see Fig. 2a. f, g, Alkaline gel electrophoresis of RNase H2 treated genomic DNA (f) shows a small increase in fragmentation for the RNASEH2A+ control clone and a more substantial increase in two independent RNASEH2A-KO clones (representative of 4 independent experiments), indicating the presence of more genome-embedded ribonucleotides compared to HeLa and parental reporter cells (g). “Control KO” cells were reported previously33,58. RFU, relative fluorescence units. h, 2 bp SSTR deletions are frequent in both RNASEH2A+ and KO cells. Mutation spectra, quantitation of indel type. Relative area of pie charts scaled to mutation rate. n, number of colonies sequenced from independent cultures. Other: complex indels or missense mutations.
Extended Data Fig. 4 | RPE1 RNase H2 null cells accumulate embedded ribonucleotides and 2–5 bp deletions across the genome. a, b, RNASEH2A and RNASEH2B KO cells (AKO, BKO, respectively) have substantially reduced cellular levels of RNase H2 subunits (a) and are deficient for RNase H2 enzyme activity (b) at the outset (ancestral) and at the end of the mutation accumulation experiment (end point). Individual data points, n = 3 technical replicates; mean ± s.d. For gel source data, see Supplementary Fig. 1. c, d, Alkaline gel electrophoresis of RNase H2 treated genomic DNA (c) shows a substantial increase in fragmentation for RNASEH2A and RNASEH2B KO clones (representative of 3 independent experiments), indicating the presence of more genome-embedded ribonucleotides compared to two WT control clones (d). Densitometry plots of c. RFU, relative fluorescence units. As RNase H2 deficiency activates the p53 pathway, experiments were performed in a TP53 knockout background. e, Only 2–5 bp deletions are significantly increased in RNase H2 null cells. Data points for acquired indel mutations in individual cell lines after 100 population doublings. Individual data points, indel counts per cell line; mean ± s.d.; P-values for two-sided Fisher’s exact test between WT (pooled counts from n = 3 independent clones) and KO (n = 2 independent clones) for one indel type vs all other indel types, after Bonferroni correction. f, g, Proportions of acquired indels in WT and KO RPE cells. After correction for indels occurring in WT, 69% of indels in RNase H2 null cells are 2–5 bp deletions. n, total indel counts. g, Quantification of 2 bp deletions by context. n, total number of 2 bp deletions. For f, g, chart areas scaled to mutation counts per line.
Extended Data Fig. 5 | ID4 occurs in RNase H2 null RPE1 cells, particularly in transcribed regions.  

a–d, Mutational spectra detected by WGS after 100 population doublings in RPE1 cells demonstrates that SSTR and SNMH/MH deletions are enriched in RNase H2 null cells. Spectra for combined RNase H2 null and wildtype cell lines (a), and individual cell lines (b). Mutational signature analysis confirms ID4 contribution in RNase H2 null (c), but not WT cells (d). e, In RNase H2 null cells, ID4 contributes significantly more to indel mutations in transcribed genomic regions ($P = 1.3 \times 10^{-29}$). Two-sided Fisher's exact test, ID4 indels vs other indels.
Extended Data Fig. 6 | ID4 mutations in RNase H2 null mouse tumours and RPE1 cells occur at a TNT motif, defining ID-TOP1. a, Mutation spectra for individual Rnaseh2b-KO mouse intestinal tumours (WGS, paired tumour–normal samples from 6 mice). b, Indel classes, detected in mouse Rnaseh2b-KO tumours. n, total indel count for 6 tumours. c, Most 2 bp deletions in these tumours occur at SSTRs and sites of single nucleotide microhomology (SNMH). n, number of 2 bp deletions. d, e, A TNT sequence motif is present at all 2 bp STR and SNMH deletions in RNase H2 null mouse tumours (d) and RPE1 cells (e). Related to Fig. 4d and Fig. 3, respectively. Sequence logo: 2-bit representation of the sequence context of 2 bp deletions. Top, all deletions, with those sequences containing a deleted adenosine (except AT/TA) reverse complemented, and deletions right-aligned. Middle, re-aligned on right-hand T. Bottom, aligned on T (STR and SNMH context only). n, number of deletions. f, Deletion sites in RNase H2 null RPE1 cells are significantly enriched for the TNT sequence motif compared to genome-wide occurrence, for all genome sequence, as well as SNMH sites. P-values, two-sided Fisher’s exact, observed vs expected. n = 98 (all; P = 8.3 x 10−14), 54 (STR; P = 0.057), 30 (SNMH; P = 0.0008) deletions.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | ID4 deletions in RNase H2 null *S. cerevisiae* occur at a TNT motif in a Top1-dependent manner. a, 2 bp deletion sites in *rnh201Δ pol2-M644G* yeast are significantly enriched for the TNT sequence motif compared to genome-wide occurrence, for all genome sequence, as well as STR sites. P-values, two-sided Fisher’s exact, observed vs expected. *n* = 94 (all; *P* = 1.0 x 10^-9), 91 (STR; *P* = 0.029), 3 (SNMH; *P* = 1) deletions. b, A TNT sequence motif is present at all 2 bp STR and SNMH deletions in *rnh201Δ pol2-M644G* yeast. Sequence logo: 2-bit representation of the sequence context of 2 bp deletions. Top, all deletions, with those sequences containing a deleted adenosine (except AT/TA) reverse complemented, and deletions aligned on right-hand T. Bottom, aligned on T (STR and SNMH context only). *n*, number of deletions. c, d, TN*T* motifs extend beyond 2 bp deletions, with enrichment above expectation for 2 bp deletions at TNT, 3 bp deletions at TNNT and 4 bp deletions at TNNNT motifs in *rnh201Δ pol2-M644G* yeast WGS. Null expectations were generated by randomly simulating deletions of 2, 3 and 4 bp (c) or 2 bp STR sequences (d) genome-wide and scoring those simulated events for TN*T* compliance. Each simulated dataset matched the count of observed mutations for the corresponding deletion class and *n* = 1,000 replicate simulated datasets were produced. The frequency distribution of TN*T* compliance in simulations is plotted as histograms, and comparison to the observed frequency of TN*T* compliance (dotted red lines) used to derive a two-tailed empirical *P*-value. e, 2 bp STR deletions have biased sequence composition. Deletions observed in *rnh201Δ pol2-M644G* yeast WGS. Genome, frequency of dinucleotides in STR sequences in mappable genome. f, Ribouridine (rU) is more common in a CrU/GrU than in an ArU/TTrU dinucleotide context. Genome-embedded ribonucleotide frequency determined by emRiboSeq<sup>86</sup>. Dotted line indicates relative rate in absence of bias (<0.25). Horizontal lines, mean; individual data points, values for *n* = 4 independent experiments<sup>85</sup>. g, h, 2 bp TNT deletions in wildtype and RNase H2 null cells are dependent on Topoisomerase 1. Mutation rates for 2 bp deletions at TNT-compliant SSTRs (g). Deletions at TNT motifs are significantly increased above expectation in WT and *rnh201Δ*, but not in *top1Δ* and *rnh201Δ top1Δ* yeast. Horizontal bars, 95% confidence intervals for odds ratio estimates (diamonds). *P*-values, two-sided Fisher’s exact after Bonferroni correction; *n* = 86, 28, 103, 19 2-bp deletions, with each deletion from an independent culture, for WT, *top1Δ*, *rnh201Δ*, *rnh201Δ top1Δ*, respectively. Null expectation, random occurrence of mutations in reporter target sequence (h).
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | TOP1-mediated mutagenesis causes increased
2–5 bp deletions in cancer. a, Of all indels, only 2–5 bp deletions are
significantly increased in CLL with biallelic RNASEH2B loss. Box, 25–75%; line, median; whiskers 5–95% with data points for values outside this range. WT (2 copies), n = 201; monallelic loss (1 copy), n = 131; biallelic loss (0 copies), n = 16 independent tumours. Indels as percentage of all variants per sample (GEL and ICGC data combined). q-values, 2-sided Mann-Whitney test with 5% FDR. b, c, In RNase H2 null CLL, 2 bp deletions predominantly occur at STR and SNMH sequences (b), and at the TNT sequence motif (c), consistent with TOP1-mediated mutagenesis. Mean ± s.e.m., percentage of all variants per sample. GEL and ICGC data combined. n = 1,711; 1,244; 443 2-bp indels identified in 201, 131, 16 biologically independent tumours, respectively. d, ID4 contribution in RNase H2 null CLL is greater in transcribed regions. Two-sided Fisher's exact test, ID4 indels vs other indels (P = 9.2 x 10^-16). e, Pan-cancer transcript expression data divided into ten expression strata for ubiquitously expressed genes (used in panel h and Fig. 5b analysis). Data points, median/maximum expression across cancer types for individual genes. Genes with similar median and maximum TPMs were considered to be ubiquitously expressed and divided into expression groups from low (1) to high (10) expression. f, Two bp deletions in cancer preferentially occur at STRs. g, ID-TOP1 deletions increase in frequency with TOP1 cleavage activity (measured by TOP1-Seq; 38). Dotted line, relative rate in lowest TOP1-seq category set to 1. Solid lines, relative deletion rate. ID-TOP1, 2–5 bp MH and SSTR deletions containing the TNN*T sequence motif. h, ID-TOP1, but not deletions in other sequence contexts, correlate with transcription. i, 2–5 bp deletions from prostate adenocarcinoma are most enriched amongst the top 10% of highly expressed prostate ‘tissue-restricted’ genes. Odds ratio (OR): number of 2–5 bp deletions in top 10% tissue restricted genes vs 2–5 bp deletions in other genes, relative to expected frequency from all other tissues. j, ID4 is not detected in the indel signature of irinotecan-treated colorectal cancers. Untreated (n = 78), treated (n = 39). k, 2–5 bp deletion frequency in cancer corresponds to TOP1 cleavage activity, in both genic and non-genic regions. Data analysed from PCAWG10, all tumours in e, h; ID4 positive tumours in g, k; Genomics England in j. In g, h and k, solid line, relative deletion rate; shading indicates 95% confidence intervals from 1,000 (g, k) or 100 (h) bootstrap replicates.
Extended Data Fig. 9 | Human germline de novo indels are enriched for ID-TOP1 deletions. 

a, Most de novo 2 bp deletions occur at SSTR, STR and SNMH sequences. b, c, A TNT sequence motif is present at the majority of 2 bp STR and SNMH deletions (b). Sequence logos: 2-bit representation of the sequence context of 2 bp deletions. Top, all deletions, with those containing A (except AT/TA) reverse complemented, and deletions right-aligned on T (where present). Bottom, STR/SNMH deletions only (c). d, TN*T motifs extend beyond 2 bp deletions, with enrichment above expectation for 2 bp deletions at TNT, 3 bp deletions at TNNT and 4 bp deletions at TNNNT motifs (P < 0.001; two-tailed empirical P-value determined for each category). Bootstrap sampling (n = 1,000) of 2, 3 and 4 bp STR/MH sequences genome-wide to derive expected frequencies of those matching TN*T motifs. Sampling was performed to match the numbers of deletions at repeats observed in the Gene4Denovo database for each category defined by repeat type, repeat unit length and total repeat length. Histograms, distribution of the number of repeats matching TN*T motifs over these samplings. Solid blue lines, kernel density estimates for these distributions. Dotted red lines, number of deletions observed in Gene4Denovo matching TN*T motifs for each category. e, ID-TOP1 correlates with germline expression level. ID-TOP1, defined as 2–5 bp MH and SSTR deletions containing the TN*T sequence motif. Shading, 95% confidence intervals from 100 bootstrap replicates.
Extended Data Fig. 10 | Topoisomerase 1 causes small deletions while protecting against topological stress. 

**a**, The canonical role of Topoisomerase 1 (TOP1) is to relieve torsional stress (sc, supercoiling) during replication and transcription. **b**, TOP1 acts by forming ssDNA nicks to release supercoils and then religates the relaxed DNA. However, TOP1 cleavage at genome-embedded ribonucleotides (frequently incorporated by replicative polymerases such as Polε), can lead to short deletions that will be most frequent at sites of torsional stress in the genome, such as occurs at highly transcribed genes. Adapted with permission from ref. 6, SpringerNature.
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 Sequencer v5.4.6; Mutation Surveyor v3.30; AIDA Image Analyzer v3.44.03S; GraphPad Prism v9.1.1; Microsoft Excel 2016; Python v3.8.5; R v3.3.1/v4.0.5/v4.1.0; SRA Toolkit v2.5.4-1; BWA-MEM v0.7.12/v0.7.15/v0.7.16; SAMBLASTER v0.1.22; GATK v3.6.0/v4.0.2.0/v4.1.9.0; Somalier v0.2.12; ISAC aligner v0.3.16.02.19; gpfBattenberg v3.2.2; ascatNgs v4.1.0; Nexus 9.0 Biodiscovery; Strelka v2.1.9.10v2.4.7/v2.8.2; SvABA v1.1.0/v1.1.3; Platypus v0.8.1; Manta v0.28.0/v1.2/v1.6.0; Canvars v1.3.1; biobambam v2.0.65; IGV v2.5.0; bcf tools v1.10.2; bedtools v2.2.30.0; samtools v1.9; iFovler (Kent source version 417); Logmaker v0.8; SigProfilerExtractor v1.1.0; SigProfilerMatrixGenerator v1.1.14/1.1.15; SigProfilerPlotting v1.1.27; scopy v1.6.3. Custom code available at https://git.ecdf.ed.ac.uk/ID-TOP1

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

RPE1 mutation accumulation experiment and mouse tumour WGS data are available from European Nucleotide Archive under accession number PRJEB48753 [https://www.ebi.ac.uk/ena/browser/view/PRJEB48753].
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.

- [x] 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-list.pdf](http://nature.com/documents/nr-reporting-summary-list.pdf).

Life sciences study design

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

| Sample size | No statistical methods were used to pre-determine sample size. Sample sizes for fluctuation assays were performed in line with standard practice in the field [doi.org/10.1385/7/59259-761-0.000]. Samples sizes for the mutation accumulation experiment were determined based on the practicality of large-scale tissue culture and time frame of the experiment (6-9 months). Sample availability of stored mouse tumours determined sample size for WGS. For CLL-GEL, CLL-IGGC, PCAWG and Gene4Denovo, sample size was determined by available data. |
| Data exclusions | CLL-GEL: samples were excluded from analysis on pre-established criteria of <50% tumour cellularity. Subsequently, 6 further cases were excluded as indel calling with Mutecc2, Platypus and/or SvABA failed. Likewise, for mouse intestinal tumours, n=3 were excluded due to low cellularity, on the basis of median SNP MAV <10%. |
| Replication | The number of times each experiment was repeated with similar results is stated in figure legends or the Methods section. Fluctuation assays were performed with a minimum of 9 independent cultures. The mutation accumulation experiment was performed on 2 independent KO clones and 3 independent WT clones. All attempts at replication were successful for the experiments described in the manuscript. |
| Randomization | Samples were allocated to groups on the basis of genotype of interest [e.g. RNase H2 status]. No randomization was performed. |
| Blinding | The investigator performing the fluctuation assays with the Hela reporter cells was blinded to the identity of the different clones. RNASEH2B genotypic assignment for CLL samples was done prior to and independent of the person who performed the indel variant count analyses. For other experiments the investigators were not blinded during data collection and analysis, as this is not standard practice in the field [enzyme assays, alkaline gels, immunoblotting, yeast fluctuation assays]. Automated colony counting was performed for fluctuation assays to avoid observer bias, and indel counts were called programmatically; therefore analyst binding was not necessary. |

Reporting for specific materials, systems and methods

[More 2021]
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

- [ ] Antibodies
- [ ] Eukaryotic cell lines
- [X] Palaeontology and archaeology
- [ ] Animals and other organisms
- [ ] Human research participants
- [ ] Clinical data
- [ ] Dual use research of concern

### Antibodies

**Antibodies used**

Sheep anti-pan-RNase H2 (raised against human recombinant RNase H2, 1:1,000); not commercially available, human RNase H2 was purified in the A.P. Jackson laboratory and antibody raised as part of a custom program by Eurogentec, and affinity purified using recombinant RNase H2; mouse anti-RNASEH2A G-10 (Santa Cruz Biotechnologies sc-515475, lot #A1416, 1:1,000); rabbit anti-GAPDH (Abcam ab9485, 1:2,000, lot #GR3380498-1); horseradish peroxidase (HRP)-linked Rabbit Anti-Sheep Immunoglobulins, (Dako, PO4163, lot #00047199, 1:2,000); Goat Anti-Mouse Immunoglobulins/HRP-linked Antibody (Dako, P0447, lot #20039214, 1:10,000); Anti-rabbit IgG, HRP-linked Antibody (Cell Signaling Technologies, 7074S, lot #29, 1:10,000).

**Validation**

RNase H2 and RNASEH2A antibodies were previously validated using knockout cell lines (doi.org/10.1038/s41598-018-0291-z); GAPDH antibody has been previously demonstrated by the manufacturer to yield a single band of the expected size and is stated on the manufacturer’s website to be cited by 2,175 publications.

### Eukaryotic cell lines

**Policy information about** [cell lines](#)

**Cell line source(s)**

HeLa cells (originally from ATCC) were a gift from G. Stewart (University of Birmingham); hTERT-RPE1 cells (originally from ATCC) were a gift from D. Durocher (The Lunenfeld–Tanenbaum Research Institute, Toronto).

**Authentication**

Cell lines were authenticated using STR DNA profiling in the labs of origin.

**Mycoplasma contamination**

All cell lines tested negative for mycoplasma contamination.

**Commonly misidentified lines**

(See [ICTA](#) register)

No commonly misidentified cell lines were used in this study.

### Animals and other organisms

**Policy information about** [studies involving animals; ARRIVE guidelines](#) recommended for reporting animal research

**Laboratory animals**

Tissues for WGS were collected from 52-week-old female Villin-Cre+ Trp53-f/f/ Rnaseh2b-f/f/ mice on a C57Bl/6j background. Mice have been described previously (doi.org/10.1053/j.gastro.2018.05.047). All mice were maintained in a specific pathogen-free facility, and the quarterly health report did not indicate the presence of pathogenic species. Mice were provided with food and water ad libitum and maintained in a 12h light–dark cycle under standard conditions (ambient temperature 20-22°C, 40-60% humidity) at Kiel University.

**Wild animals**

The study did not involve wild animals.

**Field-collected samples**

The study did not involve samples collected from the field.

**Ethics oversight**

Animal experiments were conducted with appropriate permission, in accordance with guidelines for animal care of the Christian-Albrechts-University (Kiel, Germany), in agreement with national and international laws and policies.

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

### Human research participants

**Policy information about** [studies involving human research participants](#)

**Population characteristics**

IGCC-CLL WGS cohort: age at diagnosis 18-87 years old; 91 male, 59 female. Samples were collected before administration of any treatment.

CLL Genomics England: Demographics available for 172/198 cases: age at diagnosis 38-87y (median 65); 123 male, 49 female. WGS for 174 patients prior to treatment; 24 patients at relapse or refractory to treatment.
Recruitment

ICGC-CLL: CLL patients diagnosed at the Hospital Clinic of Barcelona or at a collaborative hospital and sent to the Hospital Clinic of Barcelona for further evaluation. Only patients that fulfilled the diagnosis of CLL or MBL (Monoclonal B-cell lymphocytosis) were included.

CLL Genomics England: Patients were treatment-naive and in need of treatment according to iwCLL criteria (doi.org/10.1182/ blood-2017-09-806338) as part of their enrollment in ARTIC, AdMiRe, RIATo, FLAIR studies; apart from small subsets enrolled in CLEAR (early stage disease) and CLL129 (relapse refractory).

CRC Genomics England: Patient recruitment was organised by 13 Genomic Medicine Centres (GMCs) and their affiliated hospitals across the UK.

Ethics oversight

ICGC-CLL: All patients gave informed consent for participation in the study following the International Cancer Genome Consortium (ICGC) guidelines and the ICGC Ethics and Policy committee. The study was approved by the Research Ethics Committee of the Hospital Clinic of Barcelona.

CLL Genomics England: All patients gave written informed consent and the study was approved under the 100,000 Genomes Project Ethics Committee, East of England and South Cambridge Research Ethics Committee and the CLL Pilot Ethics approval (MREC 09/H1306/54).

CRC Genomics England: All patients gave written informed consent and the study was approved under the 100,000 Genomes Project Ethics Committee.

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