Supplementary Figure Legend

Figure S1 mSigHdp extracted signatures highly similar to COSMIC signatures

Figure S2 Detailed characterization of C\_ID5 and C\_ID8. (A) Mutational signature spectrum of C\_ID5 and C\_ID8; (B) Correlation between signature activity and age; (C) Length of deletions longer than 5bp characterized by C\_ID5 and C\_ID8; (D) The genomes strongly support the presence of C\_ID5 and C\_ID8; (E) Correlation between ID5 and ID8 activity in different tumor types reported in Alexandrov et al., 2020.

Figure S3 Comparison among COSMIC ID17, mSigHdp extracted C\_ID17 and ID\_TOP2A K743N reported by Boot et al., 2022

Figure S4 ID83 mutational signatures asymmetries in DNA regions, replication strands and transcription strands across cancer types.

(A) Enrichment of mutations in genic and intergenic regions for ID83 signatures. Each row represents one ID83 signature, and each column displays a cancer type. Signatures enriched in genic and intergenic regions are shown in circles with yellow and light blue colors, respectively. Significance of enrichment was shown as stars (Fisher’s exact test, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001).Color intensities and Circle sizes represent the odds ratio between the ratio of real mutations and the ratio of simulated mutations (See Methods).

(B) Replication strand asymmetries of ID83 signatures. Data are presented in a format similar to the one in panel (A), with green and orange colors indicating replication strand asymmetries on the lagging and leading strands, respectively.

(C) Transcription strand asymmetries of ID83 signatures. Data are presented in a format similar to the one in panel (A), with red and dark blue colors indicating transcription strand asymmetries on the transcribed and untranscribed strands, respectively.

Figure S5 Interplay between ID83 signatures and replication time.

Mutation densities of ID83 signatures per decile (y axes) are presented for early (left) to late (right) replication domains. Real mutations for signatures are shown as bars, and simulated mutations are shown as red dashed lines. As the numbers shown on top of each plot, green bars indicate signatures consistently associated with late replication timing across cancer types; yellow bars indicate signatures consistently associated with early replication timing across cancer types; purple bars indicate signatures not affected by replication timing; blue bars indicate signatures showing inconsistent trend across cancer types.

Figure S6 Characterization of MSI-associated signatures. (A) The number of SBS and indel numbers of MSS tumors, MSI tumors identified in MSI-Seq, and MSI tumors identified in both MSI-Seq and the literature. (B) Number of deletions and insertions in MSI tumors. The slope of dashed diagonal is 1. (C) Correlation between C\_ID2, H\_ID33 and H\_ID37. (D) The proportion of doublet-base deletions in tumors with ID33 presence (blue) and without ID33 presence (yellow). (E) The proportion of triplet-base deletions in tumors with ID37 presence (blue) and without ID37 presence (yellow). (F) The proportion of doublet-base deletions in tumors with H\_ID35 presence. The tumors were sorted based on their C\_ID2 activity. (G) The ROC of predictability of MSI signature activity on MSI-Seq derived MSI status (Indel83 MSI signatures only [red], Indel89 MSI signatures only (green), Indel83 and Indel89 MSI signatures together [blue]). (H) The ROC of predictability of MSI signature activity on pre-labelled MSI status (Indel83 MSI signatures only [red], Indel89 MSI signatures only (green), Indel83 and Indel89 MSI signatures together [blue]). The AUC (area under curve) was indicated in the plot.

Figure S7 Example genomes supporting the presence of H\_ID29 and InsDel29.

Figure S8 C\_ID4, H\_ID29 and mutation spectra of TOP1-TAM in vitro models. (A) C\_ID4 and H\_ID29 mutational signature; (B) Mutational spectra of Rnaseh2b-KO mice model and their cosine similarities to H\_ID29 and C\_ID4 (excluding single C/T deletion/insertion); (C) Mutational spectra of RNASEH2A KO and RNASEH2B KO model and their cosine similarities to H\_ID29 and C\_ID4 (excluding single C/T deletion/insertion); (D) Mutational spectra of pol2-M644G rnh201Δ of yeast genomes from Williams et al., 2019 and Conover et al., 2015, and their cosine similarities to H\_ID29 and C\_ID4 (excluding single C/T deletion/insertion).

Figure S9 Investigation of extended sequence context of single C/T insertions/deletions.

The sequence context of 2bp deletions at tandem repeats and 2bp deletions with single-nucleotide microhomology in (A) the top 5 samples with highest H\_ID29 activity, (B) Rnaseh2b KO mouse tumours, (C) RNase H2 null RPE1 cells, (D) RNASEH2B KO HEK293T cells and (E) the top 5 samples with highest C\_ID4 activity. In each mutation type of each model, the sequence context and the proportion of A/C/G/T on each position were displayed.

Figure S10 Contribution of Indel mutational signatures to TP53 mutations in different cancer types.Stacked bar plots illustrating the relative contributions of different indel mutational processes to mutations in genes listed in the Cancer Gene Census. The figure highlights the 5 most frequent indels for each category. For each plot, the horizontal axis represents the proportion of each mutational process, while the vertical axis displays the gene names alongside the corresponding mutation types. The specific signatures involved in each mutational process are detailed in Table S15.

Figure S11 SigProfilterExtractor result in PCAWG and HMF indel analysis. (A) Optimal solution plot generated by SigProfilterExtractor. (B) SigProfilterExtractor extracted signatures.