Figure Legend

**Figure 1 single-base substitution, doublet base substitution, and indel signatures of aristolochic acid (AA).**

A The AA indel signature according to the Indel83 classification system. The most frequent mutations are deletions of a single T (thymine) base flanked by non-thymine bases (indicated as VTV>VV) in the figure. The second most frequent deletion is a mutation of TT flanked by non T to T (VTTV>VTV). The third most frequent is a mutation from CC flanked by non C to C (indicated as DCCD>DCD).

B. The AA indel signature according to the Indel89 classification system. The deletions that are categorized as VTV>VV or VTTV>VTV in the Indel83 system are distributed to the mutation categories marked as VT{1:4}V>VT{0:3}V below the x axis, where T{1:4} denotes a string of 1, 2, 3, or 4 T and T{0:3} denotes a string of 0 to 3 T. In the Indel89 system mutations from XTY>XY, XTTY>XTY XTTTY>XTTY, and XTTTTY>XTTTY, are grouped together, but distinguished by the specific bases at X and Y. For example, in the Indel89 system the most frequent deletion type in the AA signature is denoted A[T(1,4)]A, which comprises mutations from ATA>AA, ATTA>ATA, ATTTA>ATTA, and ATTTTA>ATTTA. Similarly the most common deletion of a single C in the Indel83 system, DCCD>DCD (panel A) is allocated to “[C2]A” (DCCA>DCA), “[C2]T” (DCCT>DCT), and “C(1:5)G” (DC{1,5}G>DC{0,4}G, where C{1,5} indicates 1 to 5 Cs and C{0,4} indicates 0 to 4 Cs.

C. Single-base-substitution signature SBS22, with prominent mutation types in the signature labelled. For example, the highest peak corresponds to the proportion of CTG to CAG mutations. <we need to provide detailed labels on the X axis for this and DBS>

D. The doublet-base-substitution signature of AA.

**Figure 2 Profile of 33 Indel83 mutational signatures extracted from HMF and PCAWG**

The figure displays the profile of each of the 33 indel83 signatures, illustrating the relative contribution of various indel types and sequence contexts, including 1 bp deletions, 1 bp insertions, >1 bp deletions and insertions at repeats, and deletions with microhomology. Signatures labeled with a “C\_” prefix represent those matched to known COSMIC indel signatures, while those beginning with “H\_” denote novel signatures discovered in this study. The x-axis represents the subtype and length of indels, while the y-axis indicates the normalized proportion of each indel category within each signature.

**Figure 3** **Profile of 41 Indel89 mutational signatures extracted from HMF and PCAWG**

This figure displays the profiles of 41 indel89 mutational signatures, each representing distinct patterns of insertions and deletions across various sequence contexts. The numbering of the signatures corresponds to the previously reported Indel83 signatures, with additional sub-labels (a, b, c, d) indicating multiple Indel89 signatures derived from a single Indel83 signature. The x-axis depicts specific indel motifs and sequence features, while the y-axis represents the normalized proportion of each motif within each signature.

**Figure 4 Signature attribution of 33 Indel83 and 41 Indel89 mutational signatures**

The size of each dot represents the prevalence of a signature which indicates the proportion of genomes with exposures of the corresponding signature in the corresponding cancer type larger than 0. The color indicates the median number of exposures of the corresponding signature among samples with exposures larger than 0.

**Figure 5 Heatmap of correlation coefficients between Indel83/Indel89 indel signatures and SBS signatures.**

This figure presents a heatmap of Spearman correlation coefficients between Indel83 and Indel89 indel signatures and various SBS signatures. Signatures were excluded if all correlation coefficients were between –0.1 and 0.1. The color intensity reflects the strength of the correlation, with darker colors indicating stronger correlations. Unsupervised hierarchical clustering of both rows and columns identified six biological modules: MMR (DNA mismatch repair) defects, GI-ROS (Gastrointestinal-Reactive Oxygen Species), UV exposure, Liver, HR (DNA homologous recombination repair) defects, and Lung tobacco smoking. Each signature is annotated according to the cluster to which it belongs.

**Figure 6. Indel83 mutational signatures asymmetries in strands, replication timing and DNA regions.**

(A) Enrichment of mutations in genic and intergenic regions for Indel83 signatures. Each row represents one Indel83 signature, where n reflects the number of cancer types in which the signature was found. Signatures consistently enriched in genic and intergenic regions with p values > 0.05 (Fisher’s exact test) are shown in circles with yellow and light blue colors, respectively. Grey colors indicate inconsistent enrichments in different cancer types. Columns display the odds ratio between the ratio of real mutations and the ratio of simulated mutations (See Methods). Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of cancer types exhibiting specific enrichment in DNA regions in a signature.

(B) Replication strand asymmetries of Indel83 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 Indel83signatures. 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.

(D) Indel83signatures associated with early and late replication timing. Mutation densities of Indel83signatures 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 shown on top of each plot, green bars indicate signatures consistently associated with late replication timing across cancer types (e.g., C\_ID3); yellow bars indicate signatures consistently associated with early replication timing across cancer types(e.g., C\_ID17); purple bars indicate signatures showing inconsistent trend across cancer types(e.g., H\_ID34).

**Figure 7 MSI-associated Indel83 and Indel89 signatures**

(A-E) The mutational signature profile of 12 exclusively MSI-associated signatures, including 6 Indel83 signatures (with C\_ID7 reported in COSMIC) and 6 Indel89 signatures. (F) The activity of 7 MSI-associated signatures in MSS and MSI tumors identified in the literature or by MSI-Seq. MSI-H: Microsatellite Instability High; Non-MSI-H: Microsatellite Instability Low. (G) The proportion of sequence that removed in the MSI associated sigantures that dominated with deletions. The vertical axis represents the proportion of different deletion patterns, while the horizontal axis indicates the sequence context of each deletion event.

**Figure 8 Characterization of two novel signatures associated with TOP1-TAM**

(A) The mutational signature of TOP1-TAM (H\_ID29, InsDel29), and a previously identified mutational signature C\_ID4, and its Indel89 representation InsDel4a; (B) RNASEH2B expression in wildtype cells and three RNASEH2B KO clones and the Indel83 and Indel89 mutation spectra of one RNASEH2B KO clone. In Indel83 visualization, the INS:T:1:5+ and DEL:T:1:5+ were not displayed for a better view of the other channels; (C) The sequence logo of two-bit representation of the sequence context of 2 bp deletions at tandem repeats and 2bp deletions with single-nucleotide microhomology in 5 samples with the highest H\_ID29 activity or the highest C\_ID4 activity. (D) The signature activity in the transcribed region and the untranscribed region of tumors and in vitro models including HEK293T, mouse, RPE1 and RPE1-WT.

**Figure 9. Contribution of Indel mutational signatures to cancer gene mutations**

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 30 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. (A) Top 30 gene–deletion events in Indel83; (B) Top 30 gene–deletion events in Indel89; (C) Top 30 gene–insertion events in Indel83; (D) Top 30 gene–insertion events in Indel89. The specific signatures involved in each mutational process are detailed in Table S15.