Figure Legend

**Figure 1 The single-base substitution and doublet base substitution signatures of aristolochic acid (AA), and its indel signature in both the “ID83” and “ID89” <Maybe we should call these the “Indel83” and “Indel89” to avoid similarity to the identifiers of individual signature> systems of classifying indel mutations.** A. 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> B. The doublet-base-substitution signature of AA. C. The indel signature of AA represented using the Indel83 classification of indels. The most frequent mutations are deletions of a T (thymine) base flanked by non-thymine bases, indicated by N <change to XX> in the figure. The second most frequent deletion is a mutation of TT flanked by non T to T. The third most frequent is a mutation from CC flanked by non C to C. D. The indel signature of AA represented using the Indel89 classification of indels. The proportions of mutation categorized as XTY🡪XY or XTTY🡪XTY in the Indel83 classification are distributed to multiple types in the Indel89 classification. In this classification mutations from XTY🡪XY, XTTY🡪XTY XTTTY 🡪 XTTY, and XTTTTY 🡪 XTTTY, are group together, but distinguished by the specific bases at X and Y. For example, the most frequent mutation type in the classification is denoted A[T(1,4)]A, which comprise mutations from ATA🡪AA, ATTA->ATA, ATTTA🡪ATTA, ATTTTA🡪ATTTA.

**<From here down I renumbered the figures>**

**Figure 2 Repertoire of ID mutational signatures extracted from HMF and PCAWG**

(A) Schematic representation of the workflow of mutational signature analysis. De novo signature extraction was performed in all genomes, high TMB tumors and each of 25 cancer types. The signatures were collected from these 27 runs. The similar signatures from different runs were merged and kept. After matching with COSMIC v3.4 signatures, 18 COSMIC signatures and 15 novel signatures were identified, and used for the following analysis, including signature attribution, etiology inference, validation using in-vitro system and contribution to cancer genes. (B) 18 COSMIC signatures. The mSigHdp signatures with >0.85 cosine similarity with COSMIC 3.4 signatures. (C) 15 novel signatures. The signatures not similar to COSMIC v3.4 signatures and not be reconstructed by COSMIC v3.4 signatures.

**Figure 3 Signature attribution of 33 ID signatures**

(A) mSigAct derived signature assignment of 32 ID 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 4 Biological links indicated by the analysis of correlation between SBS and ID signatures.**

(A) Heatmap of spearman correlation coefficients between SBS and ID signatures. The coefficients were not displayed on the figure if the absolute value is less than 0.10 or the p-value of spearman correlation is larger than 0.05. Several modules were identified by hierarchical clustering based on spearman correlation coefficients: (B) tobacco smoking module; (C) cell replication module; (D) GI tract and platinum treatment module and (E) dMMR module. The shades of color and the size of dots indicate the value of spearman correlation coefficients.

**Figure 5 Characterization of ID signatures on their extended sequence context and contribution to cancer genes.**

(A) The extended sequence context characterization of DEL:C:1:0 of H\_ID24 and C\_ID9. For each plot corresponding to each ID signature, the deletion base was centered at the middle of the plot. The proportion of each nucleotide was derived for ±10nt from the deletion site, represented by red (Tyrosine), green (Adenine), blue (Cytosine) and black (Guanine). (B) The extended sequence context characterization of INS:C:1:0 of H\_ID27 and H\_ID28. The insertion cite is between ‘-1’ and ‘1’. (C) The contribution of ID signatures to most frequent mutated cancer genes. The signatures were annotated with their corresponding potential etiologies.

**Figure 6 Four novel signatures associated with MSI.**

(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) The mutational signature profile of five exclusively MSI-associated signatures, C\_ID7 (reported in COSMIC), H\_ID33, H\_ID34, H\_ID37 and H\_ID38. (D) The coefficients of spearman correlation among the activity of 7 MSI-associated signatures. (E) The activity of 7 MSI-associated signatures in MSS and MSI tumors identified in the literature or by MSI-Seq. (F) The proportion of doublet-base deletions in tumors with ID33 presence (blue) and without ID33 presence (yellow). (G) The proportion of triplet-base deletions in tumors with ID37 presence (blue) and without ID37 presence (yellow). (H) The proportion of doublet-base deletions in tumors with H\_ID35 presence. The tumors were sorted based on their C\_ID2 activity. (I) The ROC of predictability of MSI signature activity on pre-labelled MSI status (blue) and MSI-Seq MSI status (red).

**Figure 6 Characterization of a novel signature associated with TOP1-TAM**

(A) The mutational signature of TOP1-TAM (H\_ID29); (B) The mutational spectra of the 5 samples with the highest proportion of H\_ID29 activity; (C) RNASEH2B expression in wildtype cells and three RNASEH2B KO clones; (D) The mutation spectra of three RNASEH2B KO clones. The INS:T:1:5+ and DEL:T:1:5+ were not displayed for a better view of the other channels; (E) The proportion of deletion types of 2bp deletions in the 5 genomes of (B). (F) The proportion of deletion types of 2bp deletions in the 3 HEK293T clones of (D). (G) 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. (H) 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 7 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 8 Signatures associated with clinical characteristics and contribution to the cancer gene mutation**

(A) The correlation of the signature activity of C\_ID5, C\_ID9, C\_ID10 and H\_ID25 with the age of patients. (B) The enrichment evaluated by a Fisher's exact tests within each cancer type. Only the enrichment with statistical significance (p <0.05) was indicated in the figure. (C) Stacked bar plots showing the contributions of indel mutational processes to the genes listed in Cancer Gene Census. The 30 indels with the highest frequency were shown.