Figure Legend

Figure 1 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 and each of 25 cancer types. The signatures were collected from these 26 runs. The similar signatures from different runs were merged and kept, while the signatures only found once were considered as low-confidence-signatures and not used in the later analysis. After matching with COSMIC v3.4 signatures, 17 COSMIC signatures and 14 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) 17 COSMIC signatures. The mSigHdp signatures with >0.85 cosine similarity with COSMIC 3.4 signatures. (C) 14 novel signatures. The signatures not similar to COSMIC v3.4 signatures and/or not be reconstructed by COSMIC v3.4 signatures.

Figure 2 Signature attribution of 31 ID signatures

(A) mSigAct derived signature assignment of 31 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 3 Biological links indeicated by the analysis of correlation between SBS and ID signatures.

(A) mSigAct derived signature assignment of 31 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 3 Biological links indeicated by the analysis of correlation between SBS and ID signatures.

(A) mSigAct derived signature assignment of 31 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 3 Biological links indeicated by the analysis of correlation between SBS and ID signatures.

(A) mSigAct derived signature assignment of 31 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 3 Biological links indeicated by the analysis of correlation between SBS and ID signatures.

(A) Heatmap of Spearman correlation coefficient between SBS and ID signatures. mSigAct derived signature assignment of 31 ID signatures. The color represents the spearman coefficient, and the coefficient was labelled if the absolute value of the coefficient is larger than 0.05. Some signatures were not shown because no strong correlation was shown. A complete figure was provided in Figure SX. Signatures with strong correlations were clustered using hierarchical clustering on Spearman correlation coefficient: (B) The tobacco smoking module consists of SBS4, SBS92 and C\_ID3; (C) The dHR (defective homologous recombination) module consists of SBS3, SBS8, C\_ID6 and C\_ID9; (D) The GI tract and platinum treatment module consists of SBS17, SBS35, SBS88, SBS93 and C\_ID14; (E) The dMMR (defective mismatch repair) module consists of SBS6, SBS44, C\_ID7 and H\_ID33.

Figure 4 RNASEH2B deficiency signature identified from genomes and in-vitro system

(A) H\_ID29 (RNASEH2B deficiency) ID signature. H\_ID29 has predominant peaks describing 2bp deletions from 2 repeat units and 2bp deletions from microhomology, as well as 2bp deletions from 1 repeat or 3 repeat units. (B) The mutational spectra of the 5 genomes with the highest H\_ID29 activity. (C) The mutational spectra of the three RNASEH2B deficient HEK293T cell line clones. (D) The proportion of deletion types of 2bp deletions in the 5 genomes of (B). (E) The proportion of deletion types of 2bp deletions in the 3 HEK293T clones of (C).

Figure 5 Signatures associated with MSI and gender

(A) The signature pattern of 3 signatures associated with MSI status. The peaks with unique characteristics for each signatures were highlighted. (B) The signature activity of C\_ID7, H\_ID33 and H\_ID37 in MSI and MSS genomes of HMF and PCAWG cohort. (C) The proportion of deletion types derived from the 5 genomes with the highest activity of H\_ID33 (left) or H\_ID37 (right). (D) The forest plot of odds ratio and 95% CI derived from the Fisher’s exact test on the enrichment of signature presence in male or female. The signature name, Fisher’s exact test p-value, the proportion of the corresponding signatures were labelled aside of the plot.

Figure 6 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.