Mutational signatures of small insertions and deletions in 7,000 tumors

Mo Liu1,2,3\*#, Mini Huang4\*, Qi Zheng1, Arnoud Boot2,3, Shenli Zhang2,3, Szh-Chi Ho2,3, Runtian Yao4, Ying Yang1, Runxi Shen5, Steven G. Rozen2,3,6,7#

1 Sino-French Hoffmann Institute, School of Basic Medical Sciences, Guangzhou Medical University, Guangzhou, Guangdong 511436, China

2 Centre for Computational Biology, Duke–NUS Medical School, 169857 Singapore

3 Programme in Cancer and Stem Cell Biology, Duke University–National University of Singapore Medical School (Duke–NUS Medical School), 169857 Singapore

4 Molecular Cancer Research Center, School of Medicine, Shenzhen Campus of Sun Yat-Sen University, Sun Yat-sen University, Shenzhen, 518107, China

5 Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States

6 Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, 27710, United States

7 Graduate Institute of Biomedical Sciences, Chang Gung University, Taoyuan City, 33302, Taiwan

# Corresponding authors: E-mails: [mo.liu@gzhmu.edu.cn](mailto:mo.liu@gzhmu.edu.cn), steverozen@pm.me

# Abstract

Somatic mutations resulting from various mutational processes are key drivers of tumorigenesis. Mutational signatures, which are distinctive patterns left by mutational processes, can be identified through experimental exposures to mutagens or through computational analysis of somatic mutations from large collections of samples. In this study, we analyzed over 7,000 whole genomes from the Pan-Cancer Analysis of Whole Genomes and Hartwig Medical Foundation data sets to create a comprehensive collection of indel (small insertions and deletion) mutational signatures using two schemes for classifying indel mutations. We used an approach based on hierarchical Dirichlet processes approach to discover signatures according to each of the two indel classifications, and we elucidated the correspondences between the two classifications for both known and novel signatures We identified 15 signatures that were novel in both classifications, and we re-identified 23 of the COSMIC reference signatures. Of note, in cell-line experiments we showed that one pair of novel signatures, H\_ID29 and InsDel29, reflects transcription-associated mutagenesis by topoisomerase 1 at sites of genomically incorporated ribonucleotides. Among the novel signatures, four, H\_ID33, H\_ID34, H\_ID37, and H\_ID38, occurred in tumors with defective DNA mismatch repair, which were analyzed more extensively in this study. Notably, the prevalences of 9 signatures differed significantly by gender within particular cancer types. Examination of signature contributions to somatic mutations in cancer genes revealed that signartures associated with tobacco smoke exposure, account for nearly 80% of exonic indels in the tumor suppressor gene *TP53* in lung and liver cancers. This work has established an expanded collection of indel signatures in both indel classification schemes, validated a novel signature through functional modeling, elucidated indel signatures left by defective DNA mismatch repair, and has provided insights into biological implications through trait associations. <can we provide a web page for understanding these signatures?>

# Introduction

Somatic mutations are caused by various mutational processes and represent a driving force behind tumorigenesis and cancer development (Alexandrov et al. 2014). These mutations can result from either endogenous sources, such as 5-methylcytosine deamination or defective DNA repair mechanisms (Davies et al. 2017; Cooper et al. 2010; Grolleman et al. 2019; Boot et al. 2022), or exogenous sources, including exposure to chemical carcinogens in tobacco smoke or certain herbal medicines (Alexandrov et al. 2016; Ng et al. 2017; Dziubańska-Kusibab et al. 2020; Boot et al. 2020). Mutational signature analysis provides insights into cancer etiology, prognosis, prevention, evolution, and mutational signatures can also serve as biomarkers for mutagenic exposures

By “mutational signatures” we mean distinctive patterns of mutations left on genomes by mutagenic processes or exposures. They can be identified by exposing cultured cells, organoids, or experimental animals to suspected mutagens or perturbing DNA repair pathways and then sequencing the affected genomes (Boot et al. 2018; Huang et al. 2017; Kucab et al. 2019; Caipa Garcia et al. 2024; Riva et al. 2020). In addition, machine learning can identify mutational signatures as latent factors that explain the patterns of mutations in large collections of somatic mutation data (Alexandrov et al. 2020; 2014; Nik-Zainal et al. 2012; Degasperi et al. 2022; Chen et al. 2024; Jin et al. 2024; Koh et al. 2025). For example, data mining of upper tract urothelial cancers (UTUC) from Taiwan initially identified the aristolochic acid (AA) single-base substitution (SBS) signature (Hoang et al. 2013; Poon et al. 2013). Subsequent analysis revealed that this SBS signature was also present in bladder, kidney, and liver cancers (Ng et al. 2017; Poon et al. 2015). More recently, data mining of Chinese liver cancer genomes and experiments in cell culture showed that AA also generates double-base-substitution (DBS) signatures and, relevant to the current study, small insertion-and-deletion (indel) signatures (Figure 1) (Chen et al. 2024).

While mutational-signature research has emphasized SBSs, indel signatures also yield important insights into mutagenic mechanisms <put examples here>. Despite their mechanistic importance, indel signatures have received less attention than SBS signature, and as of COSMIC v3.4, 99 SBS signatures are catalogued, compared to only 23 Indel83 signatures.

There are two main systems for classifying indel mutations. The established system, termed “Indel83” here, classifies indels into 83 types and appears in Alexandrov et al. 2020 and on the COSMIC web site (https://cancer.sanger.ac.uk/signatures/id/, <https://cancer.sanger.ac.uk/signatures/documents/4/PCAWG7_indel_classification_2021_08_31.xlsx>, Figure 1A). The other system, “Indel89”, classifies indels into 89 types (Koh et al. 2025)(Figure 1B). In the current study, we follow the convention of designating Indel83 signatures with the prefix ID (e.g. ID23 in Figure 1A) and designating the Indel89 signatures that we extracted with the prefix InsDel (e.g. InsDel23 in Figure 1B). For signatures extracted in Koh et al. 2025 we use the identifiers beginning InD as presented in that paper. We have based the numbering of signatures on the numbering in <ref cosmic>. When a single Indel83 signature maps to several Indel89 signatures, we distinguish them by single-letter suffixes. For example, ID1 maps to Indel89 signatures InsDel1a, InsDel1b, InsDel1c, and InsDel1d. In many cases Koh et al assigned InD numbers to signatures with the same number in <cosmic reference>. However, some assignments appear to be discrepant based on analyses in the current study present below, and we believe the InsDel IDs provide better correspondence.

The Indel83 system primarily categorizes indels based on the number of bases inserted or deleted, and for single-base insertions and deletions the identity of the deleted or inserted base (conventionally shown as pyrimidines, C or T). For single-base indels, the Indel83 system also considers the number of flanking C or T residues (Figure 1A). Larger indels are further classified by their occurrence within repetitive sequences (i.e. microsatellite, simple tandem repeat). For deletions of ≥2 bases outside of repetitive sequences, the Indel83 system notes the presence of microhomology. Microhomology is a hallmark of non-homologous end-joining repair, which operates when repair by homologous recombination is not available, notably in BRCA-deficient tumors. For example, a 3-bp deletion (A**CA**|T**CA**|GG → A**CA**GG) exhibits a 2-base microhomology (CA). Microhomology can stem from error-prone non-homologous end joining.

Unlike the Indel83 classification, for deletions or insertions of a single T or C, the Indel89 classification incorporates distinctions based on the non-T (respectively, the non-C) flaking bases. However, the Indel89 classification makes less granular distinctions based on the number of Ts or Cs in homopolymers at the sites of single-base deletions (Figure 1B). For example, the Indel89 classification distinguishes between deletions of a T in which the preceding base is A, C, or G (in Figure 1B, signature InsDel23). Thus, deletions of ATA>AA or ATTA>ATA are distinguished from deletions of CTA>CA or CTTA>CTA, distinctions in the identity of the flanking base (in the example, A and C) that the Indel83 classification does not capture. At the same time, however, for deletions of a single T, the Indel89 classification groups together deletions of a single T in isolation along with deletions of a single T from repeats of 2 to 4 Ts, a distinction that Indel83 makes (Figure 1B versus Figure 1C). Furthermore, the Indel89 distinguishes 19 <correct?> types of indels of >= 2 bases, compared to the xx types distinguished by the Indel83 classification

The single base deletions associated with AA provide an example of the advantages of each classification system. In the Indel83 system, the AA signature shows that the most common deletion of a single T occurs as *V1*T*V2* > *V1V2*, where *V* indicates any base other than T, and *V1* need not be the same as *V2*. This is a distinction that Indel89’s AA signature does not capture. However, the Indel89 AA signature shows that deletions of T are usually associated with a flanking A, which one might hypothesize is related to the adenine adducts caused by AA. These adenine adducts are also thought to be responsible for the predominance of A>T single base substitutions in the AA SBS and DBS signatures (Figure 1C,D).

In this study, we analyzed somatic mutation data from over 7,000 tumor genomes across two large pan-cancer datasets: PCAWG (Pan-Cancer Analysis of Whole Genomes) (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al. 2020) and HMF (Hartwig Medical Foundation) (Priestley et al. 2019). Using hierarchical Dirichlet process and a non-negative matrix factorization approach, SigProfilerExtractor , we identified a comprehensive set of 33 Indel83 mutational signatures and 41 Indel89 mutational signatures. A signature was considered novel if it was not similar to any known Indel signature or could not be reconstructed from them (more details in Methods and Results). To systematically compare the two signature catalogs, we developed and applied a new pipeline to match Indel83 and Indel89 signatures based on tumor samples with high signature proportions and cosine similarities. We further profiled the replication timing, asymmetry between genic and intergenic regions and asymmetry between leading and lagging replication strand of each signature, providing insights into their underlying mutational processes. Experimental validation in a cells with deficient ribonucleotide excision repair showed that transcription-associated mutagenesis by topoisomerase 1 at sites of ribonucleotides incorporated in genomic DNA generates previously unreported indel signatures that we identified independently in both indel classification systems (Cho et al. 2013; Takahashi et al. 2011; Lippert et al. 2011) Additionally, four novel signatures from both indel classifications systems occurred predominantly in the HMF dataset, due to its larger representation of tumors with microsatellite instability (MSI). Together, our analyses provide an expanded and detailed landscape of both Indel83 and Indel89 mutational signatures, comprehensively contributions to key cancer genes, as well as their replication timing, replication strand bias, and genic versus intergenic distributions.

**Results**

***De novo* ID mutational signature discovery from large cohorts with mSigHdp.**

Where do we describe the Sigpro results? Although Non-negative Matrix Factorization (NMF) is widely used for in-silico signature discovery, complementary approaches based on hierarchical Dirichlet process may offer advantages. In particular, the R package mSigHdp (mutational signatures from hierarchical Dirichlet processes) had better benchmarking results on mutational signature discovery in synthetic Indel (and SBS) data (Liu et al. 2023). In addition, mSigHdp’s model directly infers a posterior distribution of the number of signatures present in a data set, while by contrast, NMF based approaches sometimes struggle with determining the number of signatures present (more details in Discussion).

In this study, we applied mSigHdp to identify mutational signatures from whole-genome somatic mutations across 7,013 tumors, including 2,780 from the PCAWG consortium and 4,233 from the Hartwig Medical Foundation collection (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020; Priestley et al., 2019). Mutational catalogs were generated using both the established Indel83 classification and the more recent Indel89 taxonomy. Notably, we modified one category from Koh et al.’s original classification, expanding the 1 bp C deletion from *C(6,9)* to *C(6,)*, as we observed 1 bp C deletions from polyC tracts as long as 10–15 bp in 853 samples within our dataset. Our de novo signature discovery followed a three-step approach:

1. Extraction of ID signatures was performed in three ways: (a) across all tumors combined, (b) across tumors with high tumor mutation burdens (TMB; see Methods for details), and (c) separately within each tumor type to detect tumor-specific rare signatures.
2. Highly similar signatures from all extractions were consolidated, and those reconstructible by other signatures were removed (Methods).
3. The resulting mSigHdp Indel83 signatures were compared to COSMIC v3.4 signatures and classified into three groups: (a) 18 signatures matching COSMIC v3.4 with cosine similarity > 0.85 (designated "C\_IDx," where x corresponds to the COSMIC ID; see Figure 2B and Figure S1), (b) signatures reconstructible as combinations of multiple COSMIC signatures (Methods), and (c) 15 novel signatures not fitting these categories, labeled "H\_IDx" starting from ID24, as COSMIC v3.4 ends at ID23 (Figure 2C). All novel signatures are supported by at least one sample, reinforcing their biological relevance (Vignettes); Indel89 signatures were named according to their corresponding Indel83 signatures (designated as InsDelx for matches to C\_IDx or H\_IDx, details in Methods, Vignettes). If multiple Indel89 signatures mapped to a single Indel83 signature, they were named InsDelx\_a, InsDelx\_b, and so forth.

Our analysis primarily focuses on groups (a) and (c). Overall, we identified 33 distinct Indel83 signatures and 41 Indel89 signatures (Figure 2 & Figure 3).

**Previously reported signatures**

The mSigHdp analysis successfully re-identified signatures similar to 18 of the 23 COSMIC (v3.4) Indel83 signatures (C\_ID1 to C\_ID23 in Figure2). The remaining five COSMIC signatures (ID15, ID16, ID20, ID21, and ID22) were not detected, as they are absent from the PCAWG dataset. The ability of mSigHdp to recover all COSMIC signatures present in PCAWG highlights its robustness for mutational signature analysis.

Since Indel89 signatures are not catalogued in COSMIC, we compared our Indel89 signatures to the 37 InD signatures reported by Koh et al., 2025. Of these, 21 were recapitulated in our analysis, while 10 were not identified—either due to being artefactual or absent from the PCAWG or HMF datasets. An additional six Koh et al. signatures could be reconstructed using our Indel89 signatures (see Methods and Table S1). Beyond the signatures reported by Koh et al., we identified 23 Indel89 signatures: 12 signatures map to COSMIC Indel83 signatures, 10 correspond to the novel Indel83 signatures, and one does not align with any Indel83 signature (Table 1).

Notably, there were nuanced differences between some COSMIC signatures and those extracted by mSigHdp, with our mSigHdp-derived signatures often providing more biologically plausible characterizations:

(1) ID9: Unlike C\_ID9 in our extraction, the COSMIC ID9 signature shows a near-absence of the DEL:1:T:5+ motif (Figure S1), despite DEL:1:T:5+ mutations being common in all tumors exhibiting ID9. Biologically, a process removing single thymine bases from polyT tracts of 1–4 bp would likely also operate on longer polyT stretches. The Indel89 classification supports this, as InsDel9 captures 1 bp T deletions from polyT sequences ranging from 1–9 bp. Similarly, we identified InsDel9 (comparable to Koh et al’s InD9b), which captures the depletion 1 bp T across polyT tracts of various lengths.

(2) ID5: The clock-like C\_ID5 signature incorporates elements from both COSMIC ID5 and ID8, despite a cosine similarity of 0.922 to COSMIC ID5 (Figure S2A,B). Although the long deletion patterns are highly similar between ID5 and ID8, they exhibit distinct preferences in deletion length: ID5 primarily features long deletions less than 10 nt, with almost no deletions longer than 30 nt, while ID8 displays a more even distribution of deletions ranging from 5 to over 30 nt (Figure S2C). Our analysis revealed no tumor samples supporting COSMIC ID5 in isolation, while we identified tumors that support C\_ID8 alone (Figure S2D). We examined PCAWG tumors with reported ID5 activity from Alexandrov et al., finding that there is a high correlation between ID5 activity and ID8 activity in most cancer types with both active ID5 and ID8 (Figure S2E). These findings suggest that the mutational process represented by ID5 is also responsible for long deletions in these contexts. Overall, our analysis indicates that C\_ID5 provides a more comprehensive view of genomic alterations rather than simply merging ID5 and ID8. Supporting this, the Indel89 signatures InsDel5a and InsDel5b show a high prevalence of long deletions and deletions within microhomologies, reinforcing the presence of these features in C\_ID5 (Figure 3).

(3) ID17: Compared to COSMIC ID17, we found that C\_ID17 signature enhanced the pattern of deletions at repeats and microhomologies, showing similarities to ID8 deletions. Boot et al. identified and validated an association between the TOP2A (Topoisomerase 2A) p.K743N mutation and ID17 (also known as ID\_TOP2A) using a yeast model. Our analysis revealed that our C\_ID17 signature demonstrates a closer resemblance to the ID\_TOP2A signature identified by Boot et al. than to COSMIC ID17 (Figure S3A, B, cosine similarity = 0.982).

**Signature activity**

We evaluated the activity of 33 Indel83 and 41 Indel89 signatures using mSigAct, a tool incorporating statistical analysis for the presence of a given signature (Jiang et al. 2024). Tumors with high TMB often exhibit large amounts of 1 bp T deletions and/or insertions in polyT sequences (DEL:T:1:5+ and/or INS:T:1:5+). These predominant peaks can obscure other signals, affecting the accuracy of signature assignment analysis. To address this, we propose a novel approach for analyzing signature assignments specialized for Indel83 spectra: first, we removed DEL:T:1:5+ and INS:T:1:5+ mutations to enhance the visibility of other peaks, resulting in Indel81 catalogs/signatures. Next, these Indel81 catalogs were reconstructed from the Indel81 signatures. After this reconstruction, DEL:T:1:5+ and INS:T:1:5+ mutations were reintroduced, and the signature assignment analysis was performed by comparing the original and reconstructed catalogs with C\_ID1 and C\_ID2. This strategy enables the extraction of more nuanced information from indel spectra that may otherwise be masked by high-frequency events.

Consistent with previous reports, C\_ID1, C\_ID2, C\_ID5, and C\_ID8 were detected in the majority of cancer types, while C\_ID3 was particularly prominent in lung and liver cancers, and C\_ID13 was enriched in skin cancers (Figure 4a). The novel signatures identified by mSigHdp generally exhibited activity in fewer cancer types compared to established COSMIC signatures, with the exception of H\_ID24 and H\_ID25, which were prevalent across a wide range of cancers (Figure 4b). Indel89 signature assignments revealed strong concordance with Indel83 signatures when a one-to-one mapping was present. When an Indel83 signature was represented by multiple Indel89 signatures (e.g., C\_ID1 was captured by InsDel1a, InsDel1b, InsDel1c, and InsDel1d), these split signatures often displayed cancer type-specific activity. For example, while C\_ID1 and C\_ID2 were detected across nearly all cancer types, InsDel1a was most prevalent in colon, prostate, and uterine cancers, whereas InsDel1c was enriched in biliary, CNS, lymphoid, and pancreatic cancers. Similarly, InsDel2a was most active in colon, esophagus, lymphoid, stomach, and uterine cancers, while InsDel2b was more enriched in myeloid malignancies (Figure 4a).

We further analyzed the correlations between our Indel signature activities and SBS signature activities from Degasperi et al. in both PCAWG and HMF samples (Table S2&S3). Correlation analysis revealed that InDel89 signatures can distinguish more nuanced mutational processes—showing distinct correlation patterns with SBS signatures—highlighting the greater resolution and specificity provided by the InDel89 taxonomy compared to InDel83. Unsupervised hierarchical clustering (see Methods) grouped mutational signatures into distinct biological modules. The APOBEC cluster (SBS2 and SBS13) showed moderate correlations with InsDel24b and InsDel5b. The homologous recombination deficiency (dHR) cluster (SBS3 and SBS8) exhibited strong associations with C\_ID6 and InsDel6 (Figure 5, “HR defects”). UV exposure-related signatures (SBS7a, C\_ID13, and InsDel13) also formed a distinct group. The “Liver” cluster highlights that H\_ID25 and InsDel5b, two signatures with clock-like mutational patterns, likely represent a clock-like process in liver cancer, as they show strong correlations with SBS12 and SBS16, which are both commonly observed in liver tumors. Notably, a strong correlation network was observed among mismatch repair (MMR) deficiency signatures, specifically linking SBS6, SBS26, SBS44, and a range of indel signatures including C\_ID2, InsDel2b, InsDel2c, C\_ID7, H\_ID33, InsDel33, H\_ID34, H\_ID37, H\_ID38, and InsDel38 (Figure 5, “MMR defects”). The Indel89 taxonomy provided finer resolution of mutational processes. The four Indel89 signatures derived from C\_ID1—InsDel1a to InsDel1d—exhibited distinct correlation patterns, illustrating the refined resolution provided by indel signature analysis. Notably, only InsDel1a closely mirrored the correlation profile of C\_ID1, indicating that it most accurately recapitulates the original signature. In contrast, InsDel1b frequently co-occurred with gastrointestinal-associated signatures such as SBS88, SBS17, and SBS93 within the “GI-ROS” (“Gastrointestinal-Reactive Oxygen Species) cluster, while InsDel1c was more specifically associated with SBS44. InsDel1d showed moderate correlations with SBS9, SBS17, and SBS28, which are predominantly characterized by T>C and T>G substitutions.

These correlation profiles also offer insights into the potential etiologies of previously uncharacterized SBS signatures. For example, SBS92 demonstrated a correlation profile highly similar to SBS4, particularly through strong associations with C\_ID3 and InsDel3, suggesting a possible link to tobacco smoking (Figure 5, “Lung Tobacco Smoking”). Within the GI-ROS cluster—encompassing signatures prevalent in gastrointestinal tumors such as SBS17, SBS18 (ROS), SBS1 (5-mC deamination which can be partially induced by ROS), SBS88 (colibactin exposure), and SBS93—we found that C\_ID14, C\_ID18, InsDel18, and H\_ID27 are more strongly correlated with signatures frequently observed in the GI tract. In contrast, C\_ID1, C\_ID2, and InsDel2a showed stronger correlations with SBS1 and SBS17, indicative of ROS-associated mutagenic processes (Figure 5, “GI-ROS”).

**Topography of Indel mutational signatures**

We evaluated the interplay between our Indel83 signatures and certain genomic topographical features. Transcription-coupled nucleotide excision repair was known to cause transcription strand asymmetries, since DNA bulky adducts on the transcribed strand will be preferentially repaired in the transcription active region across the genome (add cite). Transcription strand asymmetries were shown in 20 out of 33 Indel83 signatures (Figure 6A & S4C, Table S4). Signatures attributed to exogenous mutational processes, such as C\_ID3 (Tobacco smoking exposure), C\_ID14 (GI-platinum treatment associated) and C\_ID18 (Colibactin exposure), showed consistent transcription strand bias with mutations enriched in the transcribed strand. Another exogenous mutational signature C\_ID13, which is attributed to UV exposure, showed enrichment in the un-transcribed strand. This is due to the damage caused by UV happened on cytosine instead of guanine. In Indel83 signatures associated with defective endogenous mutational processes, enriched mutations on the un-transcribed strand was observed for C\_ID1(Slippage during DNA replication) and C\_ID5(Clock-like signature), while enriched mutations on the transcribed strand was observed for C\_ID29 (TOP1-mediated mutagenesis). Interestingly, 4 defective MMR signatures also showed transcription strand asymmetries in different directions, with two of them (H\_ID33 and H\_ID37) having enriched mutations on the un-transcribed strand, and two of them (H\_ID7 and H\_ID34) having enriched mutations on the transcribed strand. Replication strand asymmetries were observed in 18 out of 33 Indel83 signatures (Figure 6B & S4B, Table S4). Similar to transcription strand asymmetries, most signatures attributed to either exogenous mutagenic or defective endogenous mutational processes showed bias towards the replication leading or lagging strand. For example, 3 defective MMR signatures exhibited replication strand bias either on the leading strand (H\_ID34) or on the lagging strand (C\_ID7 and H\_ID33).

We also examined the mutation enrichment of Indel83 signatures in genic and intergenic regions (Figure S4A). Most signatures showed mutation bias towards intergenic regions, while 8 signatures showed enrichment in genic regions (Figure 6C). Among the 8 signatures, 5 of them were associated with defective endogenous mutational processes, including C\_ID17(TOP2A K743N mediated mutagenesis), H\_ID29(TOP1-mediated mutagenesis), and 3 defective MMR signatures (C\_ID7, H\_ID33 and H\_ID37). Another 3 signatures with mutation enriched in genic regions were of unknown etiology (C\_ID10, H\_ID30, H\_ID31).

The effect of DNA replication timing was also observed in many Indel83 signatures (Figure 6D & S5). Most signatures were consistently enriched in late-replicating regions across different cancer types, while only one signature C\_ID17 (TOP2A K743N mediated mutagenesis) showed enrichment in early-replicating regions. Four signatures were un-affected by replication timing in the majority of cancers presenting this signature, including C\_ID1 (Slippage during DNA replication), C\_ID5 (Clock-like signature), C\_ID13 (Ultraviolet light exposure), C\_ID18 (Colibactin exposure). Interestingly, Several signatures showed cancer-specific enrichment in replication timing, especially for defective MMR signatures. In the 5 defective MMR signatures, C\_ID7, C\_ID33 and C\_ID34 showed enriched mutations in early replication regions in breast cancer, while C\_ID37 and C\_ID38 showed this enrichment in colon cancer and stomach cancer, respectively.

**Novel Signatures**

**Microsatellite Instability associated signatures**

<change order of presentation, lead with better opportunity to analyze MSI signatures> Some microsatellite stable (MSS) tumors exhibit a high ratio of microsatellite instability (MSI) signature activity, likely due to strong MSI characteristics, such as elevated indel rates and single-base substitution (SBS) mutation loads, despite their MSS classification. Although MSI status was provided in the PCAWG and HMF datasets (Martínez-Jiménez et al. 2023; Bavi et al. 2020), several samples displaying MSI characteristics—such as high SBS and indel mutations alongside MSI-associated SBS signatures—were classified as MSS. To resolve this discrepancy, we updated the MSI status using MSISeq, a software tool designed to identify MSI status based on catalogs of somatic mutations (Huang et al. 2015). MSISeq identified an additional 98 MSI tumors beyond the 91 previously reported in the literature (Figure S6A). In total, we identified 189 MSI tumors with SBS mutation counts ranging from 10,839 to 2,432,617 and indel mutations ranging from 5,060 to 318,631. For subsequent analyses, we will refer to these 189 tumors as MSI tumors. Notably, these MSI tumors typically exhibit a higher prevalence of deletions compared to insertions (Figure S6B), suggesting that defective DNA mismatch repair predominantly leads to nucleotide removal rather than insertions.

Leveraging the higher prevalence of MSI tumors in our combined dataset, we identified five Indel83 signatures and their six corresponding Indel89 signatures associated with MSI: C\_ID2 (InsDel2b & InsDel2c, Figure 7A), C\_ID7 (InsDel7, Figure 7B), H\_ID33 and H\_ID37 (InsDel33, Figure 7C), H\_ID34 (InsDel34, Figure 7D), and H\_ID38 (InsDel38, Figure 7E). All these signatures showed significantly higher activity and enrichment in MSI tumors compared to MSS tumors (Figure 7F, Table S5). COSMIC v3.4 catalogs seven SBS signatures linked to mismatch repair (MMR) deficiency—SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44—which often co-occur and exhibit overlapping mutation patterns (e.g., SBS44 and SBS20 share nearly identical C>A profiles, while SBS6 and SBS15 both have a prominent CCG>CTG peak). we observed that H\_ID33, H\_ID37, and C\_ID7 all display >1 bp deletions at repeat sequences, but are associated with distinct indel types: C\_ID7 is mainly defined by single-base deletions of C or T from long homopolymer tracts. H\_ID33 predominantly reflects TT deletions from 4–5 TT repeats, while H\_ID37 is associated with TTT deletions from 3 TTT repeats (Figure S6D&E). Although H\_ID33 and H\_ID37 both describe similar patterns, H\_ID37 is found exclusively in tumors with high C\_ID2 activity, whereas H\_ID33 more commonly co-occurs with C\_ID2 (Figure S6C). Analysis of the corresponding Indel89 signatures revealed that InsDel33 captures the deletion patterns of both H\_ID33 and H\_ID37, characterized by a predominant peak at ‘L(2, ):U(1,2):R(5,9)’. Further examination showed that tumors with high H\_ID33 activity predominantly exhibit TT deletions from long repeats, H\_ID37 tumors show TTT and TTTT deletions, whereas C\_ID7 tumors are characterized by more dinucleotide deletions and longer polyT deletions (Figure 7G). Thus, despite shared features, these signatures arise from distinct mutational processes.

In contrast to the deletion-dominated signatures, H\_ID38 is mainly characterized by insertions—specifically 1 bp and 2 bp events at long repeats. This signature is associated with two scenarios: in samples with low C\_ID2 activity, H\_ID38 predominantly involves TT insertions; in samples with low, but not depleted, C\_ID2 activity, H\_ID38 shows a higher proportion of AT/TA insertions (Figure S6F). Among the five MSI-associated signatures, only H\_ID38 primarily describes insertions, further reinforcing the tendency for MSI tumors to exhibit more deletions than insertions.

Finally, we assessed the predictive performance of MSI signature activity as a biomarker for MSI status by performing AUROC analyses using both pre-labeled MSI status and MSISeq-derived status. Given the predominance of MSS tumors in our dataset, we evaluated predictive accuracy across all tumors, as well as subsets with ≥500 and ≥2000 indels. In every scenario, the AUROC exceeded 0.95, indicating that both Indel83 and Indel89 signature profiles provide highly accurate detection of MSI status (Figure S6G & H).

**A novel ID-TOP1 signature**

We identified a novel pair of mutational signatures, H\_ID29 and InsDel29, both characterized by 1–3 bp deletions from two repeats or microhomology, with strong support from both PCAWG and HMF samples (Figure 8A, Figure S7). Notably, two PCAWG samples exhibited high H\_ID29 activity: a skin melanoma genome (SP103894) with 3,772 H\_ID29 mutations, and a breast cancer genome (SP5559) with 949 H\_ID29 mutations (Figure S7B). The inclusion of additional samples enabled the detection of these rare signatures within the PCAWG dataset. Importantly, previous analyses often failed to extract ID4 and ID29 simultaneously, and frequently misclassified ID4 as ID-TOP1-TAM (Jin et al. 2024; Koh et al. 2025; Reijns et al. 2022). Here, for the first time, we identified C\_ID4 and H\_ID29, along with their corresponding Indel89 representations (InsDel4a, InsDel4b, and InsDel29), using a de novo extraction approach. Compared to C\_ID4, H\_ID29 (corresponding to ID-TOP1-TAM) lacks signals representing the removal of 1–3 bp sequences from regions with more than three repeats or microhomologies. This pattern is also reflected in the Indel89 representations: InsDel4a and InsDel4b show a higher proportion of peaks at *L(3,):U(3,):R(2,9)*, while InsDel29 is dominated by the peak at *L(2,8):U(1,2):R(2,4)*.

Because Indel89 analysis is currently limited to human genomes (Koh et al., 2025), and due to the strong resemblance between InDel29 and InDel4a/4b and their corresponding Indel83 signatures, we focused most of our subsequent analyses specifically on H\_ID29. Upon re-examining the rnh201Δ *Saccharomyces cerevisiae* genomes, we observed 2 bp deletion patterns similar to those of H\_ID29, although deletions within microhomology were depleted (Williams et al. 2019; Conover et al. 2015, Figures S8). We established an RNASEH2B deficiency model using the CRISPR/Cas9 system in the HEK293T cell line, and whole genome sequencing revealed patterns consistent with H\_ID29 and InsDel29 (Figure 8B). The primary peak predominantly represents the deletion of CT from 5’-CTCT-3’ (or AG from 5’-AGAG-3’), as indicated by the extended sequence analysis of RNASEH2B-KO cell lines and the five genomes exhibiting the highest H\_ID29 activity. The weights of each nucleotide on each position suggest a preference of CTCT(deletion from repeats) or NTNT (deletion with microhomology) sequences at deletion sites for H\_ID29, while NTNT (deletion from repeats) and CTNT (deletion from microhomology) at deletion sites for C\_ID4 (Figure 8C). Tumors exhibiting high H\_ID29 activity show deletion sequences that closely resemble those observed in RNASEH2B null HEK293T cells, as well as in Rnaseh2b knockout mouse tumors and RNaseH2-null RPE1 cells (Figure S6, Figure S9). In contrast, C\_ID4 displays a more balanced preference for deleting CT and TT within tandem repeats, with a prevalent CTNTN motif found in microhomologies (Figure 8C, Figure S9).

Collectively, our analysis presents H\_ID29 and InsDel29 as two novel mutational signatures identified through de novo extraction from cancer genomic data, suggesting its association with TOP1-dependent deletions in RNASEH2A and/or RNASEH2B deficient cells. Previous work by Reijns et al. developed RNASEH2A-deficient mammalian cell lines and Rnaseh2b-KO mouse intestinal cancer models, revealing the enrichment of 2 bp deletions from tandem repeats or microhomology (Reijns et al. 2022, Figure S9B, C). Our findings indicate that H\_ID29 more closely resembles the mutational spectra from these knockout models than ID4, with average cosine similarities of 0.945 in mouse models, 0.965 in human cell line models, and 0.947 in yeast models, compared to C\_ID4’s average cosine similarities of 0.690, 0.721, and 0.798 (Figure S8 B-D). Compared to ID4, H\_ID29 shows an almost depletion of long deletions (deletion length≥3) at repeats and microhomologies (Figure S6A). H\_ID29 contributes to more mutations in transcribed regions compared to untranscribed regions, which is also observed in RNase H2 null in vitro models (Figure 8D). The consistent observations suggest that H\_ID29 is associated with a transcription associated mutational process. Thus, H\_ID29 and InsDel29 provides a more accurate representation of the genomic footprints associated with TOP1-TAM (transcription-associated mutagenesis) during the cleavage of embedded ribonucleotides in the absence of RNASEH2A and/or RNASEH2B (S. N. Huang, Ghosh, and Pommier 2015; Sparks and Burgers 2015; Chon et al. 2009).

**Functional and Demographic Associations**

To assess the impact of mutational signatures on indel formation within cancer-related genes, we analyzed exonic regions of 581 Tier 1 genes from the Cancer Gene Census (Sondka et al., 2018). Deletions were most prevalent in ACVR2A, ARID1A, ATM, BAX, BCL11B, EBF1, ESR1, HNRNPA2B1, KAT6B, KMT2C, MECOM, NFIB, PBX1, PTPRK, QKI, RNF43, RPL22, TCF7L2, TPM4, and TP53 (Figure 9A&B). These events were primarily driven by DNA replication slippage, defective MMR, NHEJ, defective homologous recombination (HR), and tobacco smoking-associated signatures. Similar to insertions, TP53 deletions displayed diverse mutational patterns influenced by distinct processes: single-base cytosine deletions (DEL:C:1:1, [DEL(C):R2]T) were predominantly linked to tobacco smoking, while 2 bp deletions arising from tandem repeats or microhomologies were mediated by TOP1-TAM (H\_ID29, InsDel29) and TOP2A p.K743N (C\_ID17, InsDel17) signatures. Larger de novo deletions (>5 bp) were associated mainly with HR deficiency and NHEJ activity. The most frequently affected genes by insertions included ABI1, APC, AR, ARID1A, BAX, CDKN2A, CYLD, ERBB2, FBXO11, FGFR1, GRIN2A, HNF1A, HNRNPA2B1, MSH6, NF1, PBRM1, PHOX2B, TCF12, TCF7L2, and TP53. Notably, TP53 exhibited the highest frequency of various insertion types, such as single-base C/T insertions and long repeat insertions (INS:repeats:5+:1), largely attributable to signatures associated with DNA replication slippage, defective mismatch repair (MMR), TOP1-TAM, and non-homologous end joining (NHEJ) DNA repair (Figure 9 C&D).

We further explored the distribution of TP53 deletions across cancer types. While tobacco smoking signatures dominated in both deletion and insertions of single base cytosine in lung cancers, NHEJ DNA repair was the primary driver of TP53 deletions with length more than 5bp in bladder and biliary cancers (Figure S10). This divergence underscores how tissue-specific mutational processes drive the key gene mutations in different type of cancers

To explore potential gender differences in mutational signature activity, we performed two-sided Fisher’s exact tests with Benjamini–Hochberg correction (q < 0.2; female as reference, OR < 1 indicating male enrichment, Table S6). This analysis revealed nine significant gender-biased associations across tumor types, highlighting important biological distinctions in mutational processes between males and females. The most pronounced gender biases were found in the “Other” tumor category: both C\_ID19 (OR = 0.027, q = 2.67×10-5) and InsDel19a (OR = 0.118, q = 1.84×10-4) were strongly enriched in males, while C\_ID4 (OR = 3.13, q = 0.140) showed enrichment in females. Additional male-enriched signatures were observed in lung cancer (InsDel3: OR = 0.336, q = 0.051; C\_ID3: OR = 0.359, q = 0.101; both due to tobacco smoking) and skin cancer (InsDel13: OR = 0.407, q = 0.144; C\_ID13: OR = 0.408, q = 0.109; both due to ultraviolet light exposure). These findings likely reflect behavioral factors, as males are more likely to engage in tobacco smoking and experience greater sun exposure without adequate protection, resulting in higher prevalence of tobacco- and UV-associated mutational signatures. Conversely, female enrichment was identified for the skin-associated signature H\_ID24 (OR = 3.83, q = 0.101) and lymphoid C\_ID4 (OR = 2.49, q = 0.174). No other signature–cancer type combinations met the threshold of q < 0.2. These findings suggest that certain mutational processes—and the underlying biological mechanisms driving them—may be influenced by gender, contributing to observed differences in cancer etiology and progression.

**Discussion**

Leveraging a novel nonparametric Bayesian framework, we conducted mutational signature analysis on over 7,000 whole-genome tumor samples spanning 25 cancer types from the PCAWG and HMF cohorts. As the first study to apply both the Indel83 and Indel89 taxonomies to such a large dataset, we established a comprehensive catalog of 33 Indel83 and 41 Indel89 signatures, alongside a unified mapping system to facilitate cross-referencing of signatures between the two classifications. Notably, we discovered two novel indel signatures (H\_ID29 and InsDel29) associated with TOP1-TAM, and validated their biological relevance using both CRISPR/Cas9-mediated RNASEH2B knockout cell models and published data from RNaseH2-deficient systems. Additionally, we identified several indel signatures—four Indel83 and six Indel89—strongly associated with microsatellite instability (MSI), further elucidating the mutational footprints of defective MMR processes.

The recently developed Indel89 taxonomy by Koh et al. provided a new framework for classifying indels, particularly facilitating finer resolution of 1 bp T insertions and deletions in different contexts, and enabled more nuanced exploration of the genetic consequences of MMR defects. The study reported a 37 InDel signature database sourced from 4,775 genomes across 7 cancer types. However, challenges remain in connecting the nomenclature of Indel89 signatures to the established COSMIC Indel83 signature system; for instance, COSMIC ID5 is recognized as a clock-like signature, whereas Koh et al.'s InD5 likely reflects a sequencing artifact. By analyzing a larger and more diverse cancer genome cohort, we not only expanded the existing signature database but also systematically linked and unified the naming conventions of Indel83 and Indel89 signatures. This unified framework will support more consistent interpretation and application of indel mutational signatures (Vignette).

We also conducted signature extraction using SigProfilerExtractor, an NMF-based model known for its robust performance in signature analysis (Figure S11, Islam et al., 2022, Table S7). However, this method proved ineffective for our large cohort, yielding an optimal solution of K=12 but failing to identify several previously established COSMIC signatures. Notably, a recent study reanalyzed PCAWG indel genomes and discovered 25 Indel83 mutational signatures, including 9 novel signatures. Our analysis revealed that 3 of the 9 novel signatures identified by MuSiCal were also recapitulated in our findings (Figure S9, Jin et al., 2024). This limitation of SigProfilerExtractor is likely attributable to the challenges Non-negative Matrix Factorization faces in managing the high data sparsity and large sample size associated with indels. Our study underscores the effectiveness of mSigHdp for mining large datasets and demonstrates its capability to reveal novel signatures in highly sparse, low-count data.

The identification of novel mutational signatures often poses challenges in linking them to specific mutational processes, a complexity exacerbated by the diverse mutational landscapes observed across pan-cancer datasets. By integrating additional data into mutational signature analysis, we anticipate uncovering further signatures that more comprehensively characterize genomic mutational processes. Furthermore, we expect the development of mutational signatures as clinical biomarkers to enhance cancer diagnosis and treatment strategies.

**Materials and methods**

**Data source**

We considered two large pan-cancer whole genome cohorts: the PCAWG cohort which comprises 2780 whole-genome–sequenced samples; and the HMF cohort, comprising 3430 whole-genome–sequenced tumor samples. The Indel83 and Indel89 mutational spectra used for mutational signature extraction were provided in Table S8&S9. Variant calls for 2,780 WGS samples from the ICGC/TCGA (International Cancer Genome Consortium/The Cancer Genome Atlas) Pan-Cancer Analysis of Whole Genomes Consortium and clinical traits were obtained from the ICGC data portal (<https://dcc.icgc.org/releases/current/Projects/>, now the repository is retired, the data was downloaded on 9 May, 2024). Hartwig Medical Foundation through standardized procedures and request forms that can be found at https://www.hartwigmedicalfoundation.nl/en/appyling-for-data/. Clinical traits such as cancer type, age and gender of the HMF genomes were found from supplementary files of Priestley et al., 2019. These data was also provided in Table S10. The COSMIC Cancer Gene Census was used to identify known cancer driver genes (Sondka et al., 2018, downloaded from <https://cancer.sanger.ac.uk/cosmic/census?tier=1> on 9 Jun, 2024).

**Mutational signature extraction**

We used mSigHdp (v 2.1.2) for de novo mutational signature extraction analysis. When applying to all samples de novo mutational signatures were extracted using the cancer type to construct the hierarchy; when applying to genomes of each cancer type and high TMB genomes, the de novo mutational signatures were extracted with 2-layer HDP mixture models. In both scenario, we used the following parameters: seedNumber=1234, burnin=1000, bunin.multiplier=20, post.n = 200, post.space = 100, num.child.process=20, gamma.alpha=1, gamma.beta=50.

For SigProfilerExtractor, *de novo* mutational signatures were extracted from each mutational matrix using SigProfilerExtractor and default parameters (v1.1.24). NMF was performed with finding solutions between k = 10 and k = 30 signatures; each factorization was repeated 100 times. We ran MuSiCal with the following parameters: min\_n\_components=9, max\_n\_components=33, method=“mvnmf”, n\_replicates=100, max\_iter=10000, min\_iter=1000.

**Match mSigHdp signatures into COSMIC reference signatures**

The mSigHdp signatures were matched to previously identified COSMIC signatures (v3.4). We compared all de novo signatures to COSMIC signatures and categorized them into three groups: (1) known signature: if a mSigHdp signature has a cosine similarity of ≥ 0.85 with a COSMIC signature; (2) merged signatures: if a mSigHdp signatures can be reconstructed by at most 3 COSMIC signatures with a reconstructed similarity of ≥ 0.9; (3) novel signatures: the signatures do not fit into the known signatures or the merged signatures.

**Signature attribution analysis**

The Indel83 and Indel89 signature activities were attributed to each sample using a two-step approach: first, we used find\_best\_reconstruction\_QP function of SigTools R package (v1.0.7) to which provides a fast signature attribution analysis with quadratic programming optimization; second, we used the PresenceAttributeSigActivity function and default parameters in mSigAct R package (v3.0.1) to further refined the result from the previous step.

**Simulating synthetic cancer datasets**

Synthetic cancer datasets were simulated using SigProfilerSimulator (<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03772-3>).

**Annotating somatic indels based on transcribed versus un-transcribed strand**

We followed the method in (<https://doi.org/10.1016/j.celrep.2023.112930>). Briefly somatic indels were called with respect to the + strand of the reference genome and further annotated in regard to the pyrimidine base(s) of the insertion/deletion. Thus, indels with only C or T bases were annotated as + strand mutations; indels with only A or G bases were annotated as – strand mutations. The remaining indels were not included in the analysis. Next, + strand indels in protein coding genes were further subclassified as transcribed (template) if the gene’s sense strand was on the + strand of the genome, or else un-transcribed (sense). The logic was inverted for – strand indels. Indels in bidirectionally transcribed regions were ignored.

**Annotating somatic indels based on leading versus lagging replication strand**

Replication strand was determined by wavelet-smoothed replication-timing signal data that indicated both “valleys” (replication termination zones) and “peaks” (replication initiation zones) (<https://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/>). Valleys and peaks were sorted by the genomic coordinate in ascending order. In regard to + strand of the reference genome, replication timing signal were examined for consecutive stretches of the genome (from valley to peak or form peak to valley), with positive slope corresponded to leading strand regions and negative slope corresponded to lagging strand regions. Then for the - strand of the reference genome, leading regions (- slope) and lagging regions (+slope) were automatically acquired. Similar to the annotation for transcription, indels were first annotated as + or – strand mutations based on the pyrimidine bases. Next, indels were counted as being on leading strand or lagging strand based on their occupancy in a leading or lagging region.

**Detecting strand asymmetries across cancer types**

Strand asymmetry analyses were based on the assignment of signature probabilities to each individual indel mutation. Only indels with the probability greater than or equal to 0.50 to a certain ID signature were retained. For each ID signature and for all cancer types having this mutational signature, we retrieved the number of indels on each strand/region. In strand asymmetries analyses, only cancer types with at least 1,000 somatic mutations unambiguously attributed to an individual mutational signature were included.

For each strand asymmetry analyses (genic and intergenic region asymmetry, transcription strand asymmetry, replication strand asymmetry), indel mutations were split into two types and counted (genic vs intergenic mutation, leading strand vs lagging strand mutation, un-transcribed strand vs transcribed strand mutation). The two types were denoted as +/- strand mutations strand mutations in all three cases. The ratio of real somatic indels and the ratio of simulated somatic indels was calculated separately:

Odds ratio between the ratio of real somatic indels and the ratio of simulated somatic indels was calculated:

p values were calculated for the odds ratio using Fisher’s exact test. Only strand asymmetries with p value > 0.05 were considered showing strand asymmetries.

**Analyses of replication timing across cancer types**

Replication timing data were obtained from (<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1509-y>) The replication time signals were sorted in a descending order and subsequently divided into deciles. Somatic indels were counted within the corresponding deciles based on their overlap with the replication domains in the examined deciles. As with other analyses, for each individual ID signature, the reported replication timing analyses included only cancer types with at least 1,000 somatic mutations unambiguously attributed to it.

Replication timing mutation counts were generated for both real and simulated somatic indels. To classify whether the replication timing mutation density was increasing, flat, or decreasing, two linear regression models were fitted to the values of the real somatic indels count R(X1, X2, X3, … X10) and the values of simulated somatic indel counts S(X1, X2, X3, … X10), respectively.

In the above formula, replication timing denotes a vector c(1,2,3 … 10).

The replication timing trend for an ID signature was determined for both real and simulated data. An ID signature was considered to be generally unaffected by replication timing if the slope m was not statistically significant from a flat line. Otherwise, with the slope m statistically significant from a flat line, an ID signature was considered to be increasing from early to late replicating regions if the slope m > 0, and was considered to be decreasing from early to late replicating regions if the slope m < 0.

If the trends of a certain ID signature for the two dataset (real and simulated data) were different, the relative trend for real data comparing the simulated data was the final trend for the ID signature. Otherwise, if the trends of a certain ID signature for the two dataset were the same, a third multiple linear regression model was fitted.

Similarly, the final trend of an ID signature were determined by the significance of the coefficient of replication timing. If it is not significant, an ID signature was considered to be generally unaffected by replication timing. Otherwise, if the coefficient of replication timing > 0, an ID signature was considered to be increasing from early to late replicating regions; if the coefficient of replication timing < 0, an ID signature was considered to be decreasing from early to late replicating regions.

**sgRNA design and Plasmid construction**

Exon 1 human RNASEH2b gene was selected for targeting. sgRNA for double-strand breaks was designed by online software (http://tools.genome-engineering.org). The sequences of targets are sgRNA1 ACCACTAGCGGAGCCGCGA and sgRNA2 GCCGGTCATCATCCACACGG. px330A-GFP and px330-S2 plasmids were gifts from Shang Li’s laboratory followed by the published protocol (Ramlee et al. 2015). Briefly, top and bottom strand primers were phosphorylating and annealing using T4 PNK from NEB (New England Biolabs, cat M0201S). Cloning of the annealed two sgRNA inserts into px330A-GFP plasmid (sgRNA1) and px330-S2 plasmid (sgRNA2) respectively. Transforming the above-mentioned reactions into One Shot™ Stbl3™ Chemically Competent E. coli (Thermofisher scientific, Cat No C737303). Subsequently restrictive enzyme digesting of px330-S2-sgRNA2 plasmid with BsaI-HF (New England Biolabs, cat NEB #R3535) and cloning the digested fragment containing sgRNA2 into px330A-GFP-sgRNA1 plasmid to form px330A-GFP sgRNA1&sgRNA2 plasmid by the golden gate assembly. Plasmid DNA was extracted and purified by QIAprep Spin Miniprep Kit (Qiagen, Cat No. 27106) and sequenced to ensure the correct sgRNA sequences.

**Cell culture and plasmid transfection**

HEK293T cells were maintained in DMEM (Dulbecco's Modified Eagle Medium, Gibco™, Cat. No.11995065) containing 10% FBS (Gibco™ Fetal Bovine Serum, Cat. No. A5256801) and 1% Penicillin-Streptomycin (10,000 U/mL, Gibco™, Cat. No. 15140122) and incubated at 37°C incubator supplied with 5% CO2. Cells were seeded at 2× 105 cells per well of a 6-well plate and transfected with 2 μg of plasmid next day using Lipofectamine™ 3000 Transfection Reagent (Invitrogen™, Cat. No. L3000150) as per manufacturer’s recommendation.

**RNASEH2B KO cells selection by direct Cell lysis PCR and Western blot**

After 2 days of transfection, HEK293T cells were sorted to GFP-positive single cell into 96-well plate using FACSAria III (BD Biosciences). The single cell was continued to culture around 2 to 3 weeks in the 96-well plate until the colony could be visualized by eyes. Cell colonies were trypsinized by 10ul of 0.05% (1:10 dilution of 0.5% Trypsin-EDTA no phenol red, Gibco™, Cat. No.  15400054). Cell suspension was divided to half. Half of the cells were kept in culture. The rest half of cells was added to 10μl of Direct-Lyse lysis buffer (10mM Tris pH 8.0, 2.5mM EDTA, 0.2M NaCl, 0.15% SDS, 0.3% Tween-20) in PCR tube. The cells were then subjected to a series of heating and cooling to ensure complete lysis: 65 °C for 30s, 8 °C for 30s, 65 °C for 1.5min, 97 °C for 3min, 8 °C for 1min, 65 °C for 3min, 97 °C for 1min, 65 °C for 1min, and 80 °C for 10min (Joung et al. 2017). The lysates were then diluted with 40μl of water and cell lysis PCR was performed as regular PCR under the conditions: Initial denaturation, 5 min at 95 °C. Denaturation, 15 sec at 95 °C. Annealing, 15 sec at 58 °C. Extension, 15 sec at 72 °C for 28 cycles following final extension 10 min. Primers used for detected wild type RNASEH2b and knock-out RNASEH2b were RNASEH2B\_Wt\_Fwd-GCCCTGCTTCTGTGATCCTA, RNASEH2B\_Wt\_Rev-TCGCTTTGAACTACCCTTGG and RNASEH2B\_ko\_Fwd- CGCAGACCCAATCCTAGC, RNASEH2B\_ko\_Rev: TCCCTAGGCCAAATTCCTTT. Discard the cells which the PCR product only showed wild type band. Cells with completed knockout of RNASEH2b gene was confirmed by Western blot. 15 μg of whole cell lysis was used for immunoblotting (1:500 dilution of RNaseH2B Monoclonal Antibody cat. No. MA5-23523).

**Whole genome sequencing of RNASEH2B KO cell clones**

Genomic DNA of the completed knout out RNASEH2B cells were extracted using DNeasy Blood & Tissue Kit (Qiagen cat no.69506) and sent for whole genome sequencing (NovogeneAIT Singapore).

**MSI/MSS status and high/low TMB status**

For PCAWG genomes, the MSI status was evaluated by the PCAWG working group and obtained from the synapse repository (<https://www.synapse.org/#!Synapse:syn8016399>, the data was downloaded on May 2022). For HMF genomes, the MSI status was downloaded from the supplementary data of Priestley et al., 2019. The genomes with >14,000 IDs and >15,000 SBSs were labelled as high TMB tumors. The thresholds were selected based on the minimum number of mutations of the pre-defined MSI tumors. We then used MSI-seq to predict the MSI status of high TMB tumors.

**Extended sequence context analysis**

To analyze a specific signature and indel type of interest, we first identified the 5 genomes with the highest contribution of the corresponding signature activity. From these genomes, we extracted all indels of the relevant type. We then examined the nucleotide sequence within a 21-base pair window centered on each indel site (±10 nucleotides from the indel position). For each position within this window, we calculated the frequency of each nucleotide (A, T, C, and G). The logo was plotted based on the frequency matrix by seqLogo function of seqLogo R package (version 1.71.0)

**Acknowledgement**

We thank Shang Li for providing plasmids for generating CRISPR knockout clones.

**Fundings**

The work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 32500546), International Science and Technology Innovation Cooperation between Governments for National Key R&D Program Projects (2023YFE0107700), Guangzhou Municipal Science and Technology Bureau (2025A04J4304), the 111 Project (D18010).

**Author contributions**

M.L. and S.G.R. conceived the project and designed the analyses. M.L. and S.G.R. designed computational analysis. M.L., Q.Z. and Y.Y. performed mutational signature extraction, attribution and characterization computational analysis. M.H. and R.Y. characterized the topography of Indel83 mutational signatures. A.B., S.Z. and S.H. designed and performed gene editing experiments. R.S. performed the AUROC analysis. Data interpretation and manuscript write-up were carried out by M.L., M.H., and S.G.R., with contributions from all other authors. All authors reviewed and approved the final manuscript.

**Reference**

Alexandrov, Ludmil B., Young Seok Ju, Kerstin Haase, et al. 2016. ‘Mutational Signatures Associated with Tobacco Smoking in Human Cancer’. *Science* 354 (6312): 618–22. https://doi.org/10.1126/science.aag0299.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, et al. 2020. ‘The Repertoire of Mutational Signatures in Human Cancer’. *Nature* 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.

Alexandrov, Ludmil B, Serena Nik-zainal, David C Wedge, and Samuel A J R Aparicio. 2014. *Signatures of Mutational Processes in Human Cancer*. 500 (7463): 415–21. https://doi.org/10.1038/nature12477.Signatures.

Bavi, Prashant, Stephen B. Baylin, Wojciech Bazant, et al. 2020. ‘Pan-Cancer Analysis of Whole Genomes’. *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Boot, Arnoud, Mi Ni Huang, Alvin W.T. Ng, et al. 2018. ‘In-Depth Characterization of the Cisplatin Mutational Signature in Human Cell Lines and in Esophageal and Liver Tumors’. *Genome Research* 28 (5): 654–65. https://doi.org/10.1101/gr.230219.117.

Boot, Arnoud, Alvin W.T. Ng, Fui Teen Chong, et al. 2020. ‘Characterization of Colibactin-Associated Mutational Signature in an Asian Oral Squamous Cell Carcinoma and in Other Mucosal Tumor Types’. *Genome Research* 30 (6): 803–13. https://doi.org/10.1101/gr.255620.119.

Caipa Garcia, Angela L., Jill E. Kucab, Halh Al-Serori, et al. 2024. ‘Tissue Organoid Cultures Metabolize Dietary Carcinogens Proficiently and Are Effective Models for DNA Adduct Formation’. *Chemical Research in Toxicology* 37 (2): 234–47. https://doi.org/10.1021/acs.chemrestox.3c00255.

Chen, Lei, Chong Zhang, Ruidong Xue, et al. 2024. ‘Deep Whole-Genome Analysis of 494 Hepatocellular Carcinomas’. *Nature*, ahead of print, March 21. https://doi.org/10.1038/s41586-024-07054-3.

Cho, Jang-Eun, Nayun Kim, Yue C. Li, and Sue Jinks-Robertson. 2013. ‘Two Distinct Mechanisms of Topoisomerase 1-Dependent Mutagenesis in Yeast’. *DNA Repair* 12 (3): 205–11. https://doi.org/10.1016/j.dnarep.2012.12.004.

Cooper, David N, Matthew Mort, Peter D Stenson, Edward V Ball, and Nadia A Chuzhanova. 2010. *Methylation-Mediated Deamination of 5-Methylcytosine Appears to Give Rise to Mutations Causing Human Inherited Disease in CpNpG Trinucleotides, as Well as in CpG Dinucleotides*. http://www.hgmd.org.

Davies, Helen, Dominik Glodzik, Sandro Morganella, et al. 2017. ‘HRDetect Is a Predictor of BRCA1 and BRCA2 Deficiency Based on Mutational Signatures’. *Nature Medicine* 23 (4): 517–25. https://doi.org/10.1038/nm.4292.

Degasperi, Andrea, Xueqing Zou, Tauanne Dias Amarante, et al. 2022. ‘Substitution Mutational Signatures in Whole-Genome–Sequenced Cancers in the UK Population’. *Science* 376 (6591). https://doi.org/10.1126/science.abl9283.

Dziubańska-Kusibab, Paulina J., Hilmar Berger, Federica Battistini, et al. 2020. ‘Colibactin DNA-Damage Signature Indicates Mutational Impact in Colorectal Cancer’. *Nature Medicine* 26 (7): 1063–69. https://doi.org/10.1038/s41591-020-0908-2.

Grolleman, Judith E., Richarda M. de Voer, Fadwa A. Elsayed, et al. 2019. ‘Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-Tumor Phenotype’. *Cancer Cell* 35 (2): 256-266.e5. https://doi.org/10.1016/j.ccell.2018.12.011.

Hoang, Margaret L., Chung-Hsin Chen, Viktoriya S. Sidorenko, et al. 2013. ‘Mutational Signature of Aristolochic Acid Exposure as Revealed by Whole-Exome Sequencing’. *Science Translational Medicine* 5 (197). https://doi.org/10.1126/scitranslmed.3006200.

Huang, Mi Ni, Willie Yu, Wei Wei Teoh, et al. 2017. ‘Genome-Scale Mutational Signatures of Aflatoxin in Cells, Mice, and Human Tumors’. *Genome Research* 27 (9): 1475–86. https://doi.org/10.1101/gr.220038.116.

Huang, Mini, John R. McPherson, Ioana Cutcutache, Bin Tean Teh, Patrick Tan, and Steven G. Rozen. 2015. ‘MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations’. *Scientific Reports* 5 (1): 13321. https://doi.org/10.1038/srep13321.

Jiang, Nanhai, Yang Wu, and Steven G Rozen. 2024. ‘A New Approach to the Challenging Problem of Mutational Signature Attribution’. *bioRxiv*, ahead of print. https://doi.org/10.1101/2024.05.20.594967.

Jin, Hu, Doga C. Gulhan, Benedikt Geiger, et al. 2024. ‘Accurate and Sensitive Mutational Signature Analysis with MuSiCal’. *Nature Genetics* 56 (3): 541–52. https://doi.org/10.1038/s41588-024-01659-0.

Joung, Julia, Silvana Konermann, Jonathan S Gootenberg, et al. 2017. ‘Genome-Scale CRISPR-Cas9 Knockout and Transcriptional Activation Screening’. *Nature Protocols* 12 (4): 828–63. https://doi.org/10.1038/nprot.2017.016.

Koh, Gene Ching Chiek, Arjun Scott Nanda, Giuseppe Rinaldi, et al. 2025. ‘A Redefined InDel Taxonomy Provides Insights into Mutational Signatures’. *Nature Genetics*, ahead of print, April 10. https://doi.org/10.1038/s41588-025-02152-y.

Kucab, Jill E., Xueqing Zou, Sandro Morganella, et al. 2019. ‘A Compendium of Mutational Signatures of Environmental Agents’. *Cell* 177 (4): 821-836.e16. https://doi.org/10.1016/j.cell.2019.03.001.

Lippert, Malcolm J., Nayun Kim, Jang-Eun Cho, et al. 2011. ‘Role for Topoisomerase 1 in Transcription-Associated Mutagenesis in Yeast’. *Proceedings of the National Academy of Sciences* 108 (2): 698–703. https://doi.org/10.1073/pnas.1012363108.

Liu, Mo, Yang Wu, Nanhai Jiang, Arnoud Boot, and Steven G Rozen. 2023. ‘mSigHdp: Hierarchical Dirichlet Process Mixture Modeling for Mutational Signature Discovery’. *NAR Genomics and Bioinformatics* 5 (1): lqad005. https://doi.org/10.1093/nargab/lqad005.

Martínez-Jiménez, Francisco, Ali Movasati, Sascha Remy Brunner, et al. 2023. ‘Pan-Cancer Whole-Genome Comparison of Primary and Metastatic Solid Tumours’. *Nature* 618 (7964): 333–41. https://doi.org/10.1038/s41586-023-06054-z.

Ng, Alvin W T, Song Ling Poon, Mi Ni Huang, et al. 2017. *Aristolochic Acids and Their Derivatives Are Widely Implicated in Liver Cancers in Taiwan and throughout Asia*. https://www.science.org.

Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, et al. 2012. ‘Mutational Processes Molding the Genomes of 21 Breast Cancers’. *Cell* 149 (5): 979–93. https://doi.org/10.1016/j.cell.2012.04.024.

Poon, Song Ling, Mi Ni Huang, Yang Choo, et al. 2015. ‘Mutation Signatures Implicate Aristolochic Acid in Bladder Cancer Development’. *Genome Medicine* 7 (1): 38. https://doi.org/10.1186/s13073-015-0161-3.

Poon, Song Ling, See-Tong Pang, John R. McPherson, et al. 2013. ‘Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool’. *Science Translational Medicine* 5 (197). https://doi.org/10.1126/scitranslmed.3006086.

Ramlee, Muhammad Khairul, Tingdong Yan, Alice M. S. Cheung, Charles T. H. Chuah, and Shang Li. 2015. ‘High-Throughput Genotyping of CRISPR/Cas9-Mediated Mutants Using Fluorescent PCR-Capillary Gel Electrophoresis’. *Scientific Reports* 5 (1): 15587. https://doi.org/10.1038/srep15587.

Reijns, Martin A. M., David A. Parry, Thomas C. Williams, et al. 2022. ‘Signatures of TOP1 Transcription-Associated Mutagenesis in Cancer and Germline’. *Nature* 602 (7898): 623–31. https://doi.org/10.1038/s41586-022-04403-y.

Riva, Laura, Arun R. Pandiri, Yun Rose Li, et al. 2020. ‘The Mutational Signature Profile of Known and Suspected Human Carcinogens in Mice’. *Nature Genetics* 52 (11): 1189–97. https://doi.org/10.1038/s41588-020-0692-4.

Takahashi, Diane T., Guenaelle Burguiere-Slezak, Patricia Auffret Van Der Kemp, and Serge Boiteux. 2011. ‘Topoisomerase 1 Provokes the Formation of Short Deletions in Repeated Sequences upon High Transcription in *Saccharomyces Cerevisiae*’. *Proceedings of the National Academy of Sciences* 108 (2): 692–97. https://doi.org/10.1073/pnas.1012582108.