Mutation footprints of small-insertion-and-deletions from large cancer genomics data

Mo Liu1,3,4#, Qi Zheng1, Arnoud Boot2,3, Shenli Zhang2,3, Szh-Chi Ho2,3, Mini Huang5, Ying Yang1, Runxi Shen6, Steven G. Rozen2,3,4#

Affiliation

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

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

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

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

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

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

# 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 these processes, can be identified through experimental exposures or computational deconvolution of mutation catalogs. In this study, we analyzed over 7,000 whole genomes from the PCAWG (Pan-Cancer Analysis of Whole Genomes) and the HMF (Hartwig Medical Foundation) cohorts to create a comprehensive collection of ID (small insertions and deletions) mutational signatures using a hierarchical Dirichlet process-based approach. This analysis led to the identification of 15 novel signatures, in addition to the 23 currently cataloged in COSMIC. More specifically, we identified one novel signature, H\_ID29, associated with TOP1-TAM (Topoisomerase1 transcription-associated mutagenesis), using CRISPR/Cas9-induced knockout cells. Moreover, we identified four new dMMR (defective DNA mismatch repair) signatures—H\_ID33, H\_ID34, H\_ID37, and H\_ID38—characterizing short deletions or insertions in repeat units within tumors exhibiting high mutation burdens. Notably, three ID signatures demonstrated significant gender bias. Our examination of signature contributions to cancer genes revealed that C\_ID3, associated with tobacco exposure, accounts for nearly 50% of IDs in LRP1B, which is implicated in lung carcinogenesis. This work establishes an expanded collection of ID signatures, validates a novel signature through functional modeling, elucidates distinct mutational processes, and offers insights into biological implications through extended sequence investigation and trait associations. This comprehensive characterization of ID signatures from over 7,000 genomes enhances our understanding of the mutational processes shaping cancer genomes.

**Introduction**

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

Mutational signatures are distinctive patterns of mutations left on genomes by specific mutagenic processes or exposures. They can be identified through two approaches: (1) 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; M. N. Huang et al. 2017; Kucab et al. 2019; Caipa Garcia et al. 2024; Riva et al. 2020); and/or (2) using machine learning to deconvolute large-scale somatic mutation data (Alexandrov, Kim, et al. 2020; Alexandrov et al. 2014; Nik-Zainal et al. 2012; Degasperi et al. 2022; Chen et al. 2024; Jin et al. 2024). For instance, data mining of liver cancer genomes detected several types of mutational signature due to aristolochic acid exposure. These consisted of single-base-substitution (SBS), double-base-substitution (DBS), and insertion-and-deletion (ID/indels) signatures. These were further validated in cell-culture experiments (Chen et al. 2024).

While the characterization of mutational signatures has primarily concentrated on SBSs, ID signatures also offer valuable insights into mutagenic mechanisms. For instance, the tobacco smoking-associated mutational process not only includes C>A (SBS4) and CC>AA (DBS2) changes but also involves the removal of 1 bp C from polyC sequences of lengths 1-5, as indicated by ID3. However, the investigation of ID signatures has been comparatively neglected. To date, COSMIC v3.4 has collected 99 reference SBS signatures but only 23 ID signatures (<https://cancer.sanger.ac.uk/signatures/>).

<paragraph needs revisiting> The classification of ID mutational signatures typically involves the examination of several key features, including indel length, sequence context and indel type. Indel length refers to the number of nucleotides inserted or deleted, ranging from a single base pair to larger genomic fragments. Sequence context encompasses the nucleotide composition surrounding the indel site, which may provide insights into the underlying mutagenic mechanisms or sequence preferences. In this context, we do not consider complex indel events involving a combination of insertions and deletions. It considers insertions and deletions of single base pair of C or T, longer fragments from repeats or microhomologies, and results in 83 indel types (ID83).

In this study, we collected somatic mutation data from over 7,000 tumor genomes across two large pan-cancer datasets: PCAWG (Pan-Cancer Analysis of Whole Genomes) [Alexandrov, Ally, et al. 2020] and HMF (Hartwig Medical Foundation) [Priestley et al. 2019]. By systematically analyzing and classifying ID mutational signatures in these cancer genomes using a Hierarchical Dirichlet Process-based tool, we established a repertoire of 33 ID mutational signatures, including 15 novel signatures and several updated known signatures. We validated a novel ID mutational signature associated with TOP1-TAM (Topoisomerase 1-transcription-associated mutagenesis) within the context of RNASEH2B deficiency by investigating the genetic background and conducting in vitro experiments. Additionally, leveraging the higher rate of microsatellite instability (MSI) in the HMF dataset, we identified four novel ID signatures significantly associated with MSI status. Our analysis encompassed clinical characteristics, extended sequence contexts, and contributions to key cancer genes, providing a comprehensive characterization of ID mutational signatures.

**Results**

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

As Non-negative Matrix Factorization (NMF) is widely used for signature discovery analysis, the tool based on a non-parametric Bayesian approach demonstrates significant advantages. This approach allows for the automatic inference of optimal solutions and the sensitive and accurate extraction of mutational signatures from large cohorts. The development of Hierarchical Dirichlet Process (HDP) based extraction model mSigHdp allows a more sensitive and accurate extraction of ID signatures from large scales of genomics data (Liu et al. 2023). We performed *de novo* mutational signature analysis using mSigHdp on a total of 7,013 whole-genome sequencing (WGS) samples. This dataset comprises 2,780 genomes from the PCAWG dataset and 4,233 genomes from the HMF dataset. The extraction was performed in three ways: (1) all samples together, (2) samples with high tumor mutation burdens (TMBs, details in Method) and (3) analyzing each individual tumor type separately to identify tumor-type-specific rare signatures (Figure 1A).

We then consolidated highly similar signatures from all extractions and removed those that can be reconstructed by other signatures. Next, we compared our mSigHdp-extracted signatures to those in COSMIC v3.4 and categorized them into three groups: (1) previously reported signatures (matching COSMIC v3.4 with cosine similarity > 0.85), labeled "C\_IDX" (Figure 1B, Figure S1); (2) merged signatures combining multiple COSMIC v3.4 signatures; and (3) novel signatures not fitting the previous categories, labeled "H\_IDX" (Figure 1C). Notably, all signatures reported here are supported by at least one sample, ensuring their presence in our dataset (Figure S2). Our analysis focuses on groups (1) and (3), omitting merged signatures as they are explicable by known signatures from (1). In total, we identified 33 distinct mutational signatures.

**Previously report signatures**

Our analysis successfully reproduced 18 out of 23 COSMIC (v3.4) ID signatures. The remaining 5 signatures were either derived from whole-exome sequencing (WES) data (e.g., ID15 and ID16) or from studies not utilizing PCAWG or HMF data (e.g., ID20, ID21, ID22). In summary, mSigHdp's capability to identify nearly all COSMIC signatures underscores its reliability in mutational signature analysis.

Furthermore, several noteworthy differences were observed, and we believe that mSigHdp provides a more biologically reasonable analysis: (1) In contrast to the C\_ID9 identified in our extraction, the COSMIC ID9 signature exhibits a near-depletion of the INS:1:T:5+ motif. This discrepancy may arise from the prevalence of the INS:1:T:5+ peak in almost all tumors. Biologically, a mutagenic process removing a single thymine base from polyT sequences of lengths 1-4 would likely occur in longer polyT sequences as well. (2) 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 S3A,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 S3C). Our analysis revealed no tumor samples supporting COSMIC ID5 in isolation, while we identified tumors that support C\_ID8 alone (Figure S3D). 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 S3E). 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. (3) 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 S4A, B, cosine similarity = 0.982).

**Signature activity**

We evaluated the activity of our 33 mSigHdp signatures using mSigAct, a tool incorporating statistical analysis for the presence of a given signature (Jiang, Wu, and Rozen 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 indel spectra: first, we removed DEL:T:1:5+ and INS:T:1:5+ mutations to enhance the visibility of other peaks, resulting in ID81 catalogs/signatures. Next, these ID81 catalogs were reconstructed from the ID81 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 method allows for the extraction of more detailed information in indel spectra that may be obscured by the presence of DEL:T:1:5+ and INS:T:1:5+.

Consistent with previous studies, C\_ID1, C\_ID2, C\_ID5, and C\_ID8 were detected across most cancer types, with C\_ID3 showing a strong presence in lung and liver cancers and C\_ID13 prominently observed in skin cancers. The novel signatures identified by mSigHdp were generally active in fewer cancer types compared to COSMIC signatures, with the exception of H\_ID24 and H\_ID25, which were widespread across various cancers (Figure 2). We analyzed the correlations between our ID signature activities and the SBS signature activities from Degasperi et al. in PCAWG and HMF samples (Table S5). Our analysis confirmed strong correlations among C\_ID3, SBS4, and SBS92, all linked to tobacco-induced lung cancer (Spearman correlation coefficients: 0.75 between C\_ID3 and SBS4, 0.59 between C\_ID3 and SBS92, Figure 3A). Additionally, a strong correlation was observed between C\_ID13 and SBS7a, both associated with UV exposure (Spearman correlation coefficient: 0.81, Figure 3A).

Highly correlated genes were clustered into several interesting modules: we identified a module of four signatures related to cell replication: SBS1 (5mC deamination during cell replication), SBS18 (linked to reactive oxygen species), C\_ID1 and C\_ID2 (replication slippage) (Figure 3B). A dHR module was identified consisting of SBS3, SBS8 and C\_ID6: SBS3 and C\_ID6 were classified related to defective HR DNA damage repair, which suggests the potential etiology of SBS8 (Figure 3C). A correlation module was also noted, including C\_ID14, SBS35, SBS88, and SBS93 (Figure 3D). SBS88, and SBS93 are frequently observed in gastrointestinal (GI) tracts, while SBS35 is associated with platinum treatment, suggesting a possible etiology for C\_ID14 related to platinum treatment in GI tract cancers. Notably, we identified a dMMR (defective DNA mismatch repair) module comprising five signatures: SBS44, C\_ID7, H\_ID33, H\_ID37, and H\_ID38 (Figure 3E). Interestingly, only 1 out of 7 dMMR SBS signatures was strongly associated with indels, indicating a distinct mutational process underlying SBS44 compared to the other SBS signatures.

**Novel Signatures**

**Extended sequence context characterization of novel signatures**

We observed that some signatures share the same dominant peaks, prompting an investigation into whether they represent distinct mutational processes. To explore this, we examined the extended sequence contexts of samples with high activity for these signatures to better understand the preferential sequence context of the indels.

Both H\_ID24 and C\_ID9 display a similar pattern of 1 bp C deletions (DEL:C:1:0). However, analysis of their extended sequence contexts revealed that H\_ID24 preferentially deletes C from 5'TTTCX3', while C\_ID9 favors deletion from 5'XCTTT3' (Figure 4A). These findings suggest that H\_ID24 and C\_ID9 originate from distinct mutational processes: H\_ID24 preferentially removes cytosine 3' of poly-T sequences, whereas C\_ID9 removes cytosine 5' of poly-T sequences. Additionally, DEL:C:1:0 is prominent in H\_ID32, where the extended sequence surrounding DEL:C:1:0 shows a balanced ratio of A and T.

Furthermore, both H\_ID27 and C\_ID14 exhibit high levels of INS:C:1:0, with extended sequence analysis indicating that the INS:C:1:0 of these signatures preferentially occurs within poly-G sequences (Figure 4B). In addition, several HMF samples strongly support the presence of H\_ID27, leading us to propose that H\_ID27 is a variant form of C\_ID14, characterized by a lower proportion of INS:T:1:5+ (Figure S5).

H\_ID32 primarily consists of 1 bp C/T insertions and deletions in TA-rich sequences, while H\_ID26 describes T insertion sequences with a higher number of A bases (Figure 4C, D). Although H\_ID27 and H\_ID28 both display 1 bp C insertions (INS:C:1:0), they represent two distinct processes: H\_ID27 preferentially inserts a cytosine 3' of poly-A sequences, while H\_ID28 inserts a cytosine or guanine 3' of poly-G sequences. Based on these observations, we conclude that H\_ID27 and H\_ID28 arise from two distinct mutational processes rather than an over-splitting of a single process. Additionally, the primary mutation types in H\_ID28 exhibit a similar pattern in extended sequence context analysis; specifically, the insertion of repeats, along with 1 bp C and 1 bp T, tends to occur 3' of poly-G sequences (Figure 4B, E). In summary, characterizing the extended sequence contexts highlights the specific sequence preferences of mutational processes. Moreover, it serves as a critical tool for distinguishing signatures with similar dominant peaks, thereby determining whether they represent distinct mutational processes or variations of the same process.

**MSI signatures**

Some microsatellite stable (MSS) tumors exhibit a high ratio of 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, 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.). MSISeq identified an additional 98 MSI tumors beyond the 91 previously reported in the literature. 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 (Figure 5A). Notably, these MSI tumors typically exhibit a higher prevalence of deletions compared to insertions (Figure 5B), suggesting that defective DNA mismatch repair predominantly leads to nucleotide removal rather than insertions.

By leveraging the higher prevalence of MSI tumors in the aggregated dataset, we identified four additional MSI-associated ID signatures beyond COSMIC ID7: H\_ID33, H\_ID34, H\_ID37, and H\_ID38 (Figure 4C). COSMIC v3.4 lists seven single-base substitution (SBS) signatures associated with mismatch repair (MMR) deficiency: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44. These signatures frequently co-occur and exhibit overlapping mutation patterns; for example, SBS44 and SBS20 display nearly identical C>A mutation profiles, while SBS6 and SBS15 share a prominent CCG>CTG peak.

In our examination of ID signatures, we observed similar patterns: H\_ID33, H\_ID37, and C\_ID7 all exhibit >1 bp deletions at repeat sequences but are associated with distinct ID types (Figure 5C). We evaluated the relationships among these five signatures and two other replication slippage and MSI-associated signatures (C\_ID1 and C\_ID2). The four MSI signatures demonstrated high correlation with one another, suggesting they arise from associated downstream pathways of defective MMR. Conversely, C\_ID1—characterized by 1 bp T insertions into polyT sequences—showed negative correlations with the other MSI signatures. H\_ID34 primarily describes 1 bp T deletions from short T sequences and does not correlate with any other MSI signatures. The exclusivity between C\_ID1 and C\_ID2 is contrary to the high correlation observed in non-MSI-H tumors, which further suggests that C\_ID1 and C\_ID2 have different characteristics in MSI-H tumors compared to others (Figure S6A). Importantly, the five MSI-associated signatures, C\_ID7, H\_ID33, H\_ID34, H\_ID37 and H\_ID38, exhibit significantly greater activity and enrichment in MSI tumors compared to MSS tumors (Figure 5E; Table S5).

C\_ID7 is characterized mainly by single-base deletions of C or T from long C or T sequences. In contrast, H\_ID33 predominantly represents TT deletions from 4-5 TT repeats, while H\_ID37 is primarily associated with TTT deletions from 3 TTT repeats (Figure 5F). Although H\_ID33 and H\_ID37 describe similar deletion patterns, H\_ID37 occurs exclusively in high C\_ID2 tumors, whereas H\_ID33 often co-occurs with C\_ID2 (Figure S6B).

In contrast to these deletion patterns, H\_ID38 is primarily characterized by insertions—specifically 1 bp and 2 bp insertions at long repeats. This signature encompasses two main scenarios related to C\_ID2 activity: (1) in samples with depleted C\_ID2 activity, it predominantly involves the insertion of TT repeats; (2) in low C\_ID2 tumors, H\_ID38 shows a higher ratio of AT/TA insertions compared to its weaker preference in higher C\_ID2 tumors (Figure 5H). Among the five identified MSI signatures, only one describes insertion patterns; this again reflects the tendency for MSI tumors to exhibit a greater prevalence of deletions than insertions (Figure 5B).

To assess the potential of MSI signature activity and its proportions as biomarkers for detecting MSI status, we conducted an area under the receiver operating characteristic curve (AUROC) analysis comparing the MSI ratio with both pre-labeled MSI status (Figure S5C) and MSISeq-identified status (Figure 5I). The analysis yielded AUROC values exceeding 0.9 for both categories of MSI status, indicating strong predictive capability.

**A novel ID-TOP1 signature**

We identified a novel mutational signature, H\_ID29, characterized by 1-3 bp deletions from two repeats or microhomology, with strong support from both PCAWG and HMF samples (Figure 6A, B). Notably, two PCAWG samples displayed significant H\_ID29 activity: a skin melanoma genome (SP103894) contained 3,772 H\_ID29 mutations, while a breast cancer genome (SP5559) had 949 H\_ID29 mutations. Analyzing additional samples allowed for the detection of rare signatures within the PCAWG datasets.

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 S7D). 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 (Figure 6C, D). 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 (Figure 6E, F). The weights of each nucleotide on each position suggest a preference of NTNT sequences at deletion sites for both H\_ID29 and C\_ID4 (Figure 6G). All five models show consistently higher activity of H\_ID29 in transcribed regions, which suggests the transcription association of H\_ID29 (Figure 6H).

Our extended sequence analysis reveals distinct sequence contexts: H\_ID29 preferentially deletes CT/TC within tandem repeats, while a common NTNT motif is identified in microhomologies (Figure 7A). 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 RNase H2 null RPE1 cells (Figure 7B-D). 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 7E).

Collectively, our analysis presents H\_ID29 as a novel mutational signature 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 S7B, 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 S7 B-D). Compared to ID4, H\_ID29 shows an almost depletion of long deletions (deletion length≥3) at repeats and microhomologies (Figure S7A). 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 6H). The consistent observations suggest that H\_ID29 is associated with a transcription associated mutational process. Thus, H\_ID29 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).

**Preferential prevalence of ID mutational signatures in clinical characteristics**

It is of interest to determine whether mutational processes, as represented by mutational signatures, exhibit preferential enrichment relative to clinical characteristics, including cancer type, gender, and age. Our analysis identified four signatures with significant aging correlations, indicative of clock-like behavior: C\_ID5, C\_ID9, C\_ID10, and H\_ID25. In general, PCAWG genomes contribute more to these aging correlations compared to HMF genomes, as evidenced by the Spearman correlation coefficients and associated p-values between signature activity and age (Figure 8A).

To evaluate the preferential prevalence of mutational signatures in relation to gender, we performed Fisher's exact tests within each cancer type. Signature presence was defined as a 5% or greater contribution to the mutational burden within each sample. Prior to these tests, we excluded four cancer types known to exhibit strong gender biases: prostate cancer (exclusive to males), and uterine, breast, and ovarian cancers (exclusive to females). Results indicated that C\_ID3 and C\_ID13 were more prevalent in males, while C\_ID4, C\_ID10, and H\_ID35 were more common in females. The higher prevalence of C\_ID3 (associated with tobacco smoking) and C\_ID13 (associated with UV exposure) in males aligns with the observation that, statistically, males tend to have higher rates of tobacco use and greater cumulative exposure to UV radiation compared to females (Figure 8B). Notably, C\_ID19 displayed a strong preference dependent on cancer type: it was enriched in females with head cancers but enriched in males with bladder, kidney, and other cancers

**Signature attributions to cancer genes**

We examined the contribution of mutational signatures to indels in cancer genes, focusing on 581 Tier 1 genes from the Cancer Gene Census (Sondka et al. 2018). We excluded DEL:1:T:5+ and INS:1:T:5+ from our analysis, as these indels are primarily contributed by C\_ID1 and C\_ID2, and single-base thymine insertions/deletions in poly-T regions rarely have significant biological impacts. The genes most frequently affected by insertions were CAMTA1, ERBB4, FHIT, FOXP1, LPP, LRP1B, NRG1, PRDM16, PTPRT, and RUNX1. Several signatures with known causes contribute to these insertions, including DNA replication slippage, defective MMR, defective HR DNA damage repair, and UV exposure. Deletions most frequently affected CAMTA1, CUX1, ERBB4, FHIT, FOXP1, GPHN, LPP, LRP1B, NRG1, and PRDM16 (Figure 8B). These deletions are primarily caused by DNA replication slippage and defective MMR. Notably, the tobacco smoking signature (C\_ID3) contributes to nearly 50% of cytosine-deletions and thymine-insertions in LRP1B. Previous research has linked LRP1B mutations to lung cancer pathogenesis (Ding et al. 2008). Our analysis potentially uncovers the mutational processes responsible for LRP1B mutations.

**Discussion**

Using a novel nonparametric Bayesian approach, we analyzed over 7,000 whole-genome sequencing (WGS) tumor samples encompassing 25 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and Hartwig Medical Foundation (HMF) cohorts. As the first study using >7000 genomes for ID signature analysis, our study established a comprehensive collection of 33 ID mutational signatures. We identified one indel signature associated with TOP1-TAM, validating this finding via CRISPR/Cas9 system and previously published RNaseH2 null in vitro models. Additionally, we found 4 ID signatures strongly linked to microsatellite instability (MSI) status, which implement the understanding of indel footprints left my defective MMR mechanism. We also performed an extended sequence context analysis to understand more information behind the formation of mutational signatures.

We also conducted signature extraction using SigProfilerExtractor, an NMF-based model known for its robust performance in signature analysis (Figure S8, Islam et al., 2022). 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 utilized a minimum-volume NMF model, MuSiCal, to reanalyze PCAWG indel genomes and discovered 25 indel 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). When we applied MuSiCal to our datasets, it resulted in an optimal K=13, with very few overlaps with COSMIC signatures (Figure S10). In contrast, using mSigHdp, we identified 30 mutational signatures across all genomes, with 24 included in the finalized collection (Table S4). This limitation of SigProfilerExtractor is likely attributable to the challenges Non-negative Matrix Factorization faces in managing the high data sparsity 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.

As sequencing technology advances, numerous national cancer research initiatives are underway. Mutational signatures have proven valuable in predicting cancer treatment efficacy and tracing disease etiology. By integrating more data into mutational signature analysis, we anticipate discovering additional signatures that 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 mutational spectra used for mutational signature extraction were provided in Table S1. 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). Variant calls for 3417 WGS samples from the HMF cohort were obtained from xxxx. 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 S2. 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 4 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 33 ID 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.

**sgRNA design and Plasmid construction**

Exon 1 human RNASEH2b gene was selected for targeting. sgRAN for double-strand breaks was designed by online software (http://tools.genome-engineering.org). The sequences of targets are sgRNA1 ACCACTAGCGGAGCCGCGA and sgRNA2GCCGGTCATCATCCACACGG.

px330A-GFP and px330-S2 plasmids were gifts from Shang Li’s laboratory (please check the cat number with him if needed) followed by the published protocol (Ref 1). 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(ref 2). 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**

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**

**Thank Shang Li for plsmid. Funding**

Reference

Alexandrov, Ludmil B., Adrian Ally, Kathryn Alsop, Eva G. Alvarez, Fernanda Amary, Samirkumar B. Amin, Brice Aminou, et al. 2020. ‘Pan-Cancer Analysis of Whole Genomes’. *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Alexandrov, Ludmil B., Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, 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, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, 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.

Boot, Arnoud, Mi Ni Huang, Alvin W.T. Ng, Szu Chi Ho, Jing Quan Lim, Yoshiiku Kawakami, Kazuaki Chayama, Bin Tean Teh, Hidewaki Nakagawa, and Steven G. Rozen. 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, Mo Liu, Nicole Stantial, Viraj Shah, Willie Yu, Karin C. Nitiss, John L. Nitiss, Sue Jinks-Robertson, and Steven G. Rozen. 2022. ‘Recurrent Mutations in Topoisomerase IIα Cause a Previously Undescribed Mutator Phenotype in Human Cancers’. *Proceedings of the National Academy of Sciences* 119 (4): e2114024119. https://doi.org/10.1073/pnas.2114024119.

Caipa Garcia, Angela L., Jill E. Kucab, Halh Al-Serori, Rebekah S. S. Beck, Madjda Bellamri, Robert J. Turesky, John D. Groopman, 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, Mo Liu, Jian Bai, Jinxia Bao, Yin Wang, et al. 2024. ‘Deep Whole-Genome Analysis of 494 Hepatocellular Carcinomas’. *Nature*, March. https://doi.org/10.1038/s41586-024-07054-3.

Chon, Hyongi, Alex Vassilev, Melvin L. Depamphilis, Yingming Zhao, Junmei Zhang, Peter M. Burgers, Robert J. Crouch, and Susana M. Cerritelli. 2009. ‘Contributions of the Two Accessory Subunits, RNASEH2B and RNASEH2C, to the Activity and Properties of the Human RNase H2 Complex’. *Nucleic Acids Research* 37 (1): 96–110. https://doi.org/10.1093/nar/gkn913.

Conover, Hailey N, Scott A Lujan, Mary J Chapman, Deborah A Cornelio, Rabab Sharif, Jessica S Williams, Alan B Clark, Francheska Camilo, Thomas A Kunkel, and Juan Lucas Argueso. 2015. ‘Stimulation of Chromosomal Rearrangements by Ribonucleotides’. *Genetics* 201 (3): 951–61. https://doi.org/10.1534/genetics.115.181149.

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, Lucy R. Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, 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, Andrea Martinez-Martinez, Gene Ching Chiek Koh, João M.L. Dias, Laura Heskin, 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.

Ding, Li, Gad Getz, David A. Wheeler, Elaine R. Mardis, Michael D. McLellan, Kristian Cibulskis, Carrie Sougnez, et al. 2008. ‘Somatic Mutations Affect Key Pathways in Lung Adenocarcinoma’. *Nature* 455 (7216): 1069–75. https://doi.org/10.1038/nature07423.

Dziubańska-Kusibab, Paulina J., Hilmar Berger, Federica Battistini, Britta A.M. Bouwman, Amina Iftekhar, Riku Katainen, Tatiana Cajuso, 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, Maartje Nielsen, Robbert D.A. Weren, Claire Palles, Marjolijn J.L. Ligtenberg, 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.

Huang, Mi Ni, Willie Yu, Wei Wei Teoh, Maude Ardin, Apinya Jusakul, Alvin Wei Tian Ng, Arnoud Boot, 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, Shar-yin Naomi, Sanchari Ghosh, and Yves Pommier. 2015. ‘Topoisomerase I Alone Is Sufficient to Produce Short DNA Deletions and Can Also Reverse Nicks at Ribonucleotide Sites’. *Journal of Biological Chemistry* 290 (22): 14068–76. https://doi.org/10.1074/jbc.M115.653345.

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

Jin, Hu, Doga C. Gulhan, Benedikt Geiger, Daniel Ben-Isvy, David Geng, Viktor Ljungström, and Peter J. Park. 2024. ‘Accurate and Sensitive Mutational Signature Analysis with MuSiCal’. *Nature Genetics* 56 (3): 541–52. https://doi.org/10.1038/s41588-024-01659-0.

Kucab, Jill E., Xueqing Zou, Sandro Morganella, Madeleine Joel, A. Scott Nanda, Eszter Nagy, Celine Gomez, 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.

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.

Ng, Alvin W T, Song Ling Poon, Mi Ni Huang, Jing Quan Lim, Arnoud Boot, Willie Yu, Yuka Suzuki, 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, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, 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.

Priestley, Peter, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, et al. 2019. ‘Pan-Cancer Whole-Genome Analyses of Metastatic Solid Tumours’. *Nature* 575 (7781): 210–16. https://doi.org/10.1038/s41586-019-1689-y.

Reijns, Martin A. M., David A. Parry, Thomas C. Williams, Ferran Nadeu, Rebecca L. Hindshaw, Diana O. Rios Szwed, Michael D. Nicholson, 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, Alastair Droop, James Hewinson, Michael A. Quail, Vivek Iyer, 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.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. ‘The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers’. *Nature Reviews Cancer* 18 (11): 696–705. https://doi.org/10.1038/s41568-018-0060-1.

Sparks, Justin L, and Peter M Burgers. 2015. ‘Error‐free and Mutagenic Processing of Topoisomerase 1‐provoked Damage at Genomic Ribonucleotides’. *The EMBO Journal* 34 (9): 1259–69. https://doi.org/10.15252/embj.201490868.

Williams, Jessica S., Scott A. Lujan, Zhi-Xiong Zhou, Adam B. Burkholder, Alan B. Clark, David C. Fargo, and Thomas A. Kunkel. 2019. ‘Genome-Wide Mutagenesis Resulting from Topoisomerase 1-Processing of Unrepaired Ribonucleotides in DNA’. *DNA Repair* 84 (December):102641. https://doi.org/10.1016/j.dnarep.2019.102641.

Ramlee, M., Yan, T., Cheung, A. *et al.* High-throughput genotyping of CRISPR/Cas9-mediated mutants using fluorescent PCR-capillary gel electrophoresis. *Sci Rep* **5**, 15587 (2015). <https://doi.org/10.1038/srep15587>

Joung, J., Konermann, S., Gootenberg, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat Protoc* **12**, 828–863 (2017). https://doi.org/10.1038/nprot.2017.016