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

Mo Liu, Arnoud Boot, Qi Zheng, Nanhai Jiang, Willie Yu, Steven G. Rozen

From Steve: Need Abstract. Have you contacted Arnoud about RNASEH2B?

**Abstract**

Somatic mutations arising from various mutational processes represent a driving force in tumorigenesis. Mutational signatures are distinctive mutation patterns left by mutagenic processes and can be identified through experimental exposures or computational deconvolution of mutation catalogs. We analyzed over 7,000 whole genomes from the PCAWG (Pan-Cancer Analysis of Whole Genomes) and HMF (The Hartwig Medical Foundation) cohorts to establish a comprehensive collection of ID (small insertions and deletions) mutational signatures using the hierarchical Dirichlet process-based tool mSigHdp. This enabled identification of 15 novel signatures beyond the 23 currently in COSMIC. We validated one novel signature, H\_ID29, associated with RNaseH2B deficiency through CRISPR/Cas9 induced knockouts. Beyond known ID7, we identified three novel dMMR (defective DNA mismatch repair) signatures, H\_ID33, H\_ID37 and H\_ID38. These showed distinct yet overlapping 1bp deletion peaks at repeat sequences and demonstrated a strong binomial relationship with MSI status. Through analysis of extended sequence contexts, distinct mutational processes were discerned for signatures with similar peaks. Gender differences in signature prevalence were also observed. Examining signature contributions to cancer genes revealed that C\_ID3, linked to tobacco exposure, accounts for nearly 50% of indels in LRP1B, implicated in lung carcinogenesis. Our analyses establish an expanded collection of ID signatures, validate a novel signature through functional modeling, discern distinct mutational processes, and provide biological insights through genomic correlations. This comprehensive characterization of ID signatures from over 7,000 genomes advances understanding of 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 (Cooper et al., 2010; Davies et al., 2017; Grollman 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, and signatures can serve as biomarkers for mutagenic exposures (Boot et al., 2020; Davies et al., 2017; Dziubańska-Kusibab et al., 2020; Grollman 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 (<alain ballmain’s lab; is there a review?> Boot et al., 2018; Huang et al., 2017; Kucab et al., 2019); and/or (2) using machine learning to deconvolute large-scale somatic mutation data (Alexandrov et al., 2014, 2020; Degasperi et al., 2022; Nik-Zainal et al., 2012). 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) signatures. These were further validation in cell-culture experiments (Chen et al., 2024)..

While the characterization of mutational signatures has mostly focused on SBSs, ID signatures also provide important information regarding mutagenic mechanisms <and can help distinguish between mutational processes with similar SBS signatures?>, but have 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/>). <partly due to inconsistency in calling somatic indels> A recent study reported the latest collection by adding 40 novel SBS signatures and 18 novel DBS signatures from 18,640 genomes (Degasperi et al., 2022). <Is this paper really getting much traction; do we really believe the 40 new SBS signatures are novel?> This shows the need of a comprehensive analysis of ID signature extraction from large cohorts.

<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 result in 83 indel types (ID83).

In this study, we collected somatic mutation data from >7000 tumor genomes from two large pan-cancer collections: PCAWG [Pan-Cancer Analysis of Whole Genomes (Aaltonen et al., 2020)] and HMF [Hartwig Medical Foundation (Priestley et al., 2019)]. By systematically analyzing and classifying indels mutational signatures in these cancer genomes using a Hierarchical Dirichlet Process based tool, we established a repertoire of 31 indel mutational signatures including 14 novel signatures and some update of known signatures. By investigating the genetic background and validation in the in-vitro experimental system, we validate a novel indel mutational signature caused by RNASEH2B deficiency. <double check the next sentence; are the new signatures really only in the HMF data?> In addition, taking the advantage of higher MSI (microsatellite instability) rate in the HMF dataset, we found 2 new ID signatures significantly associated with defective DNA mismatch repair deficiency.

**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 indel signatures from large scales of genomics data(Liu et al., 2023). We performed *de novo* mutational signature analysis using mSigHdp on a total of 7013 whole-genome sequencing (WGS) samples. This dataset comprises 2780 genomes from the PCAWG cohort and 4233 genomes from the HMF cohort. The extraction was performed in three ways: (1) aggregating all samples together, (2) MSI tumors (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 the ones 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). Our analysis concentrates on groups (1) and (3), omitting merged signatures as they are explicable by known signatures. In total, we identified 32 ID mutational signatures.

**Previously report signatures**

Our analysis successfully reproduced 16 out of 23 COSMIC (v3.4) ID signatures. The remaining 7 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), potentially exhibiting different mutational backgrounds. Notably, the signal associated with ID10 was recapitulated in H\_ID25 (Figure S1), likely due to insufficient samples strongly supporting its presence, suggesting that ID10 extraction lacks robust biological support. In summary, the ability of mSigHdp to identify nearly all COSMIC signatures demonstrates 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) mSigHdp C\_ID5 signature incorporates elements of both COSMIC ID5 and ID8, despite having a 0.94 cosine similarity to COSMIC ID5. This can be attributed to the co-occurrence of these signatures in tumors and their shared correlation with aging. We found no tumor samples supporting COSMIC ID5 alone, justifying the merger of these signatures. In our analysis, while neither C\_ID5 nor C\_ID8 showed correlations with patient age, their sum strongly correlated with patient age (Figure S2A, Spearman coefficient = 0.12, p-value = 0.0038 in HMF, Spearman coefficient = 0.18, p-value = 0.0036 in PCAWG). (3) Compared to COSMIC ID17, mSigHdp 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\_ID7 signature demonstrates a closer resemblance to the ID\_TOP2A signature identified by Boot et al. than to COSMIC ID17 (Figure S2B, C, cosine similarity = 0.982).

**Signature activity**

We evaluated the activity of our 32 mSigHdp signatures using mSigAct, a tool that outperforms others in signature attribution analysis (Jiang et al., 2024). Because MSI tumors usually carries large amounts of indels, the majority of which are 1bp T deletions on polyTs and/or 1bp C deletion on polyCs (DEL:T:1:5+ and/or DEL:C:1:5+). The predominant peaks affect the accuracy of signature assignment analysis because other signals are hidden. We therefore proposed a new way to perform signature assignment analysis for MSI tumors: we firstly removed the 1bp T deletions and 1bp C deletions to make other peaks more obvious, as well as the DEL:T:1:5+ and DEL:C:1:5+ deletions from the mutational signatures, resulting in ID81 catalogs/signatures. Then these ID81 catalogs were reconstructed by the ID81 signatures. After this, we put the DEL:T:1:5+ and DEL:C:1:5+ back and reconstructed the tumors, then perform the signature assignment analysis of the difference between catalogs and reconstructed catalogs with C\_ID1 and C\_ID2. In this way, we can reveal more detailed information in MSI tumors that cannot be observed when DEL:T:1:5+ and DEL:C:1:5+ are present.

Consistent with previous studies, C\_ID1, C\_ID2, C\_ID5, and C\_ID8 were observed across most cancer types, with C\_ID3 showing a strong presence in lung cancers and C\_ID13 showing a strong presence in skin cancers. The novel signatures identified by mSigHdp were generally active in fewer cancer types compared to COSMIC signatures, except H\_ID24 and H\_ID25 which were widespread across cancer types (Figure 2). We analyzed correlations between our ID signature assignments and the SBS signature assignments by Degasperi et al. in PCAWG and HMF samples (Figure S3). Our analysis confirmed strong correlations between C\_ID3, SBS4, and SBS92 exposures, all linked to tobacco-induced lung cancer (Spearman correlation coefficients: 0.74 between C\_ID3 and SBS4, 0.62 between C\_ID3 and SBS92, Figure 3A & B). We observed a strong correlation between C\_ID13 and SBS7a, both associated with UV exposure (Spearman correlation coefficient: 0.85, Figure 3A). C\_ID9 (unknown etiology based on COSMIC v3.4) was identified in a module of homologous recombination (HR) deficiency signatures, including C\_ID6, SBS3, and SBS8. This suggests a potential association between C\_ID9 and HR deficiency (Figure 3C). We identified a correlation module including C\_ID14, SBS93, SBS17, SBS88, and SBS35 (Figure 3D). SBS17, SBS88, and SBS93 are frequently observed in gastrointestinal (GI) tracts, while SBS35 is associated with platinum treatment. This correlation suggests a possible etiology for C\_ID14: platinum treatment in GI tract cancers.

**Novel Signatures**

**RNaseH2B deficiency preferentially removes 2bp nucleotides on genome**

A novel signature consisting of 1-3bp deletions from 2 repeats or microhomology was found in samples from HMF (Figure 4A). The major peak exclusively describes the deletion of CT from 5’-CTCT-3’ (or AG from 5’-AGAG-3’), suggested by the extended sequence context of 5 samples with the highest H\_ID29 activity, and two of these samples are from PCAWG cohort (Figure 4B). Interestingly, in addition to HMF samples, we also observed two PCAWG samples with mutant RNASEH2B carrying high H\_ID29: a skin melanoma genome SP103894 was detected with 3,772 H\_ID29 mutations, and a breast cancer genome SP5559 was detected with 949 H\_ID29 mutations (Figure 4D). The inclusion of more samples enables the discovery of rare signatures from PCAWG datasets.

We revisited the rnh201Δ *Saccharomyces cerevisiae* genomes and found highly similar patterns of 2bp deletions in the yeast genomes as H\_ID29 (). We established the RNASEH2B deficiency model by CRISPR/Cas9 system in HEK293T cell line. The whole genome sequencing shows highly similar patterns. Reijns et al., established RNASEH2A deficient mammal cell lines and proposed the 2bp deletions within ID4 are results from TOP1 (topoisomerase1)-mediated deletions via RNASEH2A deficiency (). Taken together, for the first time we found H\_ID29 by de novo extraction from cancer genomics data, and we concluded that H\_ID29 is a novel ID signature associated with TOP1-dependent deletions in RNASEH2A and/or RNASEH2B deficiency cells.

RNaseH2 is a key mammalian genome surveillance enzyme required for ribonucleotide removal. It is composed of a single catalytic subunit (A) and two non-catalytic subunits (B and C) and specifically degrades the RNA of RNA:DNA hybrids (Chon et al., 2009).

(need more details here, e.g., what position of RNASEH2B, what cell line?)

**MSI signatures**

Leveraging the higher proportion of microsatellite instability (MSI) tumors in the Hartwig Medical Foundation (HMF) dataset, we identified additional MSI-associated ID signatures beyond COSMIC ID7: H\_ID33, H\_ID37 and H\_ID38 (Figure 5A). COSMIC v3.4 lists seven SBS signatures associated with mismatch repair (MMR) deficiency: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44. These signatures often co-occur and show overlapping peaks. For instance, SBS44 and SBS20 have nearly identical C>A mutation patterns, while SBS6 and SBS15 share a predominant CCG>CTG peak. We observed similar patterns in ID signatures, with H\_ID33, H\_ID37, and C\_ID7 all showing >1bp deletions at repeat sequences, but they preferentially characterize different ID types.

C\_ID7 characterizes mainly 1bp C or T deletions from long C or T sequences, H\_ID33 mainly characterizes TT deletions from 4-5 TT repeats, while H\_ID37 mainly characterizes TTT deletions from 3 TTT repeats (Figure 5C). Different from the recapitulation of deletions, H\_ID38 describes predominantly insertions including 1bp and 2 bp insertions at long repeats, and it mainly consists of two scenarios: when a sample carries mainly insertions, the insertion mainly inserts TT repeats; when a sample carries more balanced deletions and insertions, the insertion inserts more types of dinucleotides. These MSI signatures show a strong binomial distribution of exposures in tumors: they have high activity in MSI tumors, while extremely low activity in MSS tumors (Figure 5B).

We then explore whether the high activity of MSI signatures can serve as a MSI status biomarker. We also found some samples with strong MSI signatures activity but previously were labelled as MSS, suggesting the potential of MSI signatures as a MSI status biomarker. Five genes were involved in DNA mismatch repair system: MSH2, MSH3, MSH6, MLH1, PMS2. We further checked whether the genes are mutated in samples with high MSI signatures.

**Extended sequence context characterization of novel signatures**

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

DEL:C:1:0

Both H\_ID24 and C\_ID9 exhibit 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 C deletion from 5'XCTTT3' (Figure 6A). These findings suggest that H\_ID24 and C\_ID9 arise 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 also prominent in H\_ID32, where the extended sequence surrounding DEL:C:1:0 shows a balanced ratio between A and T.

H\_ID27 and part of H\_ID28 both display 1 bp C insertions (INS:C:1:0), but they characterize 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 result from two distinct mutational processes rather than an over-splitting of a single process. Furthermore, we noted that both H\_ID27 and C\_ID14 exhibit high levels of INS:C:1:0. Extended sequence analysis indicated that the INS:C:1:0 of these two signatures preferentially occur within poly-G sequences (Figure ). Several HMF samples strongly support the presence of H\_ID27, leading us to propose that H\_ID27 is a variant form of C\_ID14 which is composed of more DEL:T:1:0. We also found that the primary mutation types in H\_ID28 share 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.

**Gender comparison**

We investigated whether mutational processes, as represented by mutational signatures, exhibit gender-specific patterns. We firstly exclude the samples with gender characteristic including prostate cancer (only in males), uterus cancer, breast cancer and ovary cancer (only in females). To assess gender-specific prevalence of mutational signatures, we employed Fisher's Exact Test. From a total of 5,000 patients with available gender data, we identified 7 signatures demonstrating significant gender-specific associations: 6 signatures (C\_ID3, C\_ID5, C\_ID8, C\_ID12, C\_ID13, H\_ID30) showed a significant prevalence in male patients. Conversely, only 1 signature (H\_ID25) was more commonly observed in female patients (Figure 5D). Some of the observations can be explained biologically, for example, C\_ID3 is associated with tobacco smoking which has a higher proportion of male patients.

**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. 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 (Figure 6C). Deletions most frequently affected CAMTA1, CUX1, ERBB4, FHIT, FOXP1, GPHN, LPP, LRP1B, NRG1, and PRDM16 (Figure 6C). 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. 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 31 insertion-deletion (ID) mutational signatures. We identified one indel signature associated with RNaseH2B deficiency, validating this finding via CRISPR/Cas9 system. Additionally, we found three ID signatures strongly linked to microsatellite instability (MSI) status, which implement the understanding of indel footprints left my defective MMR mechanism. This study, for the first time, performs an extended sequence context analysis to understand more information behind the formation of mutational signatures.

We attempted signature extraction using SigProfilerExtractor which is a NMF based extraction model with the outstanding performance in signature extraction analysis (Islam et al., 2022). However, we found it ineffective for this large cohort (optimal solution K=12): it was not able to identify novel signatures and previously identified COSMIC signatures. This is probably due to the limitations of Non-negative Matrix Factorization in handling the high data sparsity characteristic of indels (Supplementary figure). Jin et al. identified 25 indel signatures using MuSiCal, a minimum-volume NMF model (Jin et al., 2024). Our study confirmed two MuSiCal signatures using a different approach (Figure SX). Moreover, our study added more novel signatures because of integrating HMF data. Our study demonstrates the efficacy of mSigHdp in mining large datasets and reveals how this alternative approach can uncover novel signatures in highly sparse, low-count data.

In addition to HEK293T, we also generated RNASEH2B deficiency genotype in HeLa cell lines. But we didn’t observe H\_ID29, whereas we found H\_ID29 in HeLa RNASEH2A KO clones. This suggests that some additional mechanisms were added on the formation of H\_ID29, e.g., DNA repair deficiency. It is known that HEK293T has defective mismatch repair, which may also play a role in the formation of H\_ID29.

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 3417 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).

**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 samples of each cancer type, 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=100.

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.

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 ≥ 0.85; (3) novel signatures: the signatures do not fit into the known signatures or the merged signatures.

Signature attribution analysis

The 31 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.

Cell line culture and RNaseH2B CRIPSR

Need help here

MSI/MSS 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>). For HMF genomes, the MSI status was downloaded from the supplementary data of Priestley et al., 2019.

Gender enrichment by Fisher’s exact test

To evaluate the presence of mutational signatures in male and female, we used Fisher's Exact Test to determine the statistical significance of signature enrichment by gender. We quantified the frequency of the presence of each signature (exposure > 0) in both groups and applied the test to assess associations. A p-value threshold of 0.05 was established to indicate significant enrichment.

Extended sequence context

To analyze a specific signature and indel type of interest, we first identified the 10 genomes with the highest exposure to the corresponding signature. 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).

**Acknowledgement**

**Funding**

Reference

Aaltonen, L. A., Abascal, F., Abeshouse, A., Aburatani, H., Adams, D. J., Agrawal, N., Ahn, K. S., Ahn, S.-M., Aikata, H., Akbani, R., Akdemir, K. C., Al-Ahmadie, H., Al-Sedairy, S. T., Al-Shahrour, F., Alawi, M., Albert, M., Aldape, K., Alexandrov, L. B., Ally, A., … von Mering, C. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., & Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, *354*(6312), 618–622. https://doi.org/10.1126/science.aag0299

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Boutros, P., … Yu, W. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-zainal, S., Wedge, D. C., & Aparicio, S. A. J. R. (2014). *Signatures of mutational processes in human cancer*. *500*(7463), 415–421. https://doi.org/10.1038/nature12477.Signatures

Boot, A., Huang, M. N., Ng, A. W. T., Ho, S. C., Lim, J. Q., Kawakami, Y., Chayama, K., Teh, B. T., Nakagawa, H., & Rozen, S. G. (2018). In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Research*, *28*(5), 654–665. https://doi.org/10.1101/gr.230219.117

Boot, A., Ng, A. W. T., Chong, F. T., Ho, S. C., Yu, W., Tan, D. S. W., Iyer, N. G., & Roze, S. G. (2020). Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Research*, *30*(6), 803–813. https://doi.org/10.1101/gr.255620.119

Chen, L., Zhang, C., Xue, R., Liu, M., Bai, J., Bao, J., Wang, Y., Jiang, N., Li, Z., Wang, W., Wang, R., Zheng, B., Yang, A., Hu, J., Liu, K., Shen, S., Zhang, Y., Bai, M., Wang, Y., … Wang, H. (2024). Deep whole-genome analysis of 494 hepatocellular carcinomas. *Nature*. https://doi.org/10.1038/s41586-024-07054-3

Chon, H., Vassilev, A., Depamphilis, M. L., Zhao, Y., Zhang, J., Burgers, P. M., Crouch, R. J., & Cerritelli, S. M. (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

Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V, & Chuzhanova, N. A. (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, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A. M., Simpson, P. T., King, T. A., Raine, K., Eyfjord, J. E., Kong, G., Borg, Å., Birney, E., Stunnenberg, H. G., Van De Vijver, M. J., … Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, *23*(4), 517–525. https://doi.org/10.1038/nm.4292

Degasperi, A., Zou, X., Amarante, T. D., Martinez-Martinez, A., Koh, G. C. C., Dias, J. M. L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V. Y. W., Nanda, A. S., Bernstein, A., Momen, S. E., Young, J., Perez-Gil, D., Memari, Y., Badja, C., Shooter, S., Czarnecki, J., … Nik-Zainal, S. (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, P. J., Berger, H., Battistini, F., Bouwman, B. A. M., Iftekhar, A., Katainen, R., Cajuso, T., Crosetto, N., Orozco, M., Aaltonen, L. A., & Meyer, T. F. (2020). Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nature Medicine*, *26*(7), 1063–1069. https://doi.org/10.1038/s41591-020-0908-2

Grolleman, J. E., de Voer, R. M., Elsayed, F. A., Nielsen, M., Weren, R. D. A., Palles, C., Ligtenberg, M. J. L., Vos, J. R., ten Broeke, S. W., de Miranda, N. F. C. C., Kuiper, R. A., Kamping, E. J., Jansen, E. A. M., Vink-Börger, M. E., Popp, I., Lang, A., Spier, I., Hüneburg, R., James, P. A., … Kuiper, R. P. (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, M. N., Yu, W., Teoh, W. W., Ardin, M., Jusakul, A., Ng, A. W. T., Boot, A., Abedi-Ardekani, B., Villar, S., Myint, S. S., Othman, R., Poon, S. L., Heguy, A., Olivier, M., Hollstein, M., Tan, P., Teh, B. T., Sabapathy, K., Zavadil, J., & Rozen, S. G. (2017). Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Research*, *27*(9), 1475–1486. https://doi.org/10.1101/gr.220038.116

Islam, S. M. A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., He, Y., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Steele, C. D., Otlu, B., … Alexandrov, L. B. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics*, *2*(11). https://doi.org/10.1016/j.xgen.2022.100179

Jin, H., Gulhan, D. C., Geiger, B., Ben-Isvy, D., Geng, D., Ljungström, V., & Park, P. J. (2024). Accurate and sensitive mutational signature analysis with MuSiCal. *Nature Genetics*, *56*(3), 541–552. https://doi.org/10.1038/s41588-024-01659-0

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., Arlt, V. M., Phillips, D. H., & Nik-Zainal, S. (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, M., Boot, A., Ng, A. W. T., Gordân, R., & Rozen, S. G. (2021). Mutational processes in cancer preferentially affect binding of particular transcription factors. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-82910-0

Liu, M., Wu, Y., Jiang, N., Boot, A., & Rozen, S. G. (2023). MSigHdp: Hierarchical Dirichlet process mixture modeling for mutational signature discovery. *NAR Genomics and Bioinformatics*, *5*(1). https://doi.org/10.1093/nargab/lqad005

Ng, A. W. T., Poon, S. L., Huang, M. N., Lim, J. Q., Boot, A., Yu, W., Suzuki, Y., Thangaraju, S., Ng, C. C. Y., Tan, P., Pang, S.-T., Huang, H.-Y., Yu, M.-C., Lee, P.-H., Hsieh, S.-Y., Chang, A. Y., Bin, †, Teh, T., Steven, †, & Rozen, G. (2017). *Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia*. https://www.science.org

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–993. https://doi.org/10.1016/j.cell.2012.04.024

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., … Cuppen, E. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, *575*(7781), 210–216. https://doi.org/10.1038/s41586-019-1689-y

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. In *Nature Reviews Cancer* (Vol. 18, Issue 11, pp. 696–705). Nature Publishing Group. https://doi.org/10.1038/s41568-018-0060-1