Geographic variation of mutagenic exposures in kidney cancer genomes

Supplementary Information Inventory

This file contains the following:

SUPPLEMENTARY NOTE

Extraction of *de novo* mutational signatures with SigProfilerExtractor

Extractions were performed for single base substitutions (SBSs), doublet base substitutions (DBSs), and small insertions and deletions (IDs; **Supplementary Figs 1-3**). SBS extractions were performed using both SBS-288 and SBS-1536 contexts. These two contexts extend the SBS-96 classification in two independent ways. SBS-288 by considering the SBS-96 contexts on transcribed and untranscribed strands of protein coding genes as well as by including mutations on intergenic non-transcribed DNA. SBS-1536 is formed of a pentanucleotide context formed of the two flanking bases on both the 5' and 3' of the mutated base. Although using different information to extract mutational signatures, the two extractions were largely concordant (**Supplementary Table 1**) with two observed differences:

1) SBS-1536 was able to extract an additional signature (SBS1536C/SBS_C, **Supplementary Fig. 1**). Visual inspection of this signature showed a similarity to SBS1536H/SBS_H, a signature which was driven by a single hypermutator case (**Supplementary Fig. 4**). On this basis, we hypothesised that these two signatures were more distinct at the SBS-1536 level than at the SBS-288 level which would lead to a failure to extract them separately in the SBS-288 extraction. To test this, we

performed an SBS-288 extraction with the hypermutator case removed, which was able to extract a signature corresponding to SBS1536C/SBS_C, thus supporting its existence.

2) SBS-1536 extracted an additional 'flat' signature (SBS1536E/SBS_E), in addition to three extracted in both SBS-288 and SBS-1536 format (**Supplementary Fig. 1**). Flat signatures are termed so because they lack distinct peaks in specific contexts but rather have variable peaks spread across all substitution types. The consequence of this is that they are very difficult to accurately distinguish between, and again, the difference between the two extractions is very likely to be due to the signature being more distinct at the SBS-1536 level compared to the SBS-288.

For our dataset, the SBS-1536 format was selected for further analysis due to its ability to extract additional signatures. In principle, whether SBS-288 or SBS-1536 is more effective is likely to vary between datasets, depending on which signatures are present and the amount of overlap between substitution types.

Extraction of *de novo* signatures with mSigHdp

To ensure that the mutational signature landscape was fully reflected in the SigProfilerExtractor results, extraction of mutational signatures was also performed with mSigHdp. In contrast to SigProfilerExtraction, which utilises nonnegative matrix factorization, mSigHdp leverages a hierarchical Dirichlet process.[1] However, unlike SigProfilerExtractor, mSigHdp has only been benchmarked for use with SBS-96 and ID-83 contexts which prevents a comprehensive direct comparison. mSigHdp extracted 11 SBS signatures and 6 ID signatures (**Supplementary Figs 5-6**). The 11 SBS signatures were largely concordant with the SBS-288 signatures (**Supplementary Table 1**), and the same differences compared to SBS-1536 described above were observed. The exception to this was the second Aristolochic acid signature hdp11, which only had a cosine similarity of 0.75 and 0.77 to SBS1536I (SBS_I)

and SBS288J respectively. Despite the differences in the signature extracted, the mSigHdp results support the existence of a second SBS Aristolochic acid signature. Given that this second Aristolochic acid signature has not been identified in prior studies which have been largely focused on SBS-96 contexts, it is likely that the extended contexts are important for distinguishing between the two signatures. mSigHdp extracted 6 ID-83 signatures, one less than SigProfilerExtractor. Overall, the results were similar, with mSigHdp extracting ID_A and ID_B as a single signature (hdp2) explaining the difference in the number of signatures extracted.

Decomposition to reference signatures

Decomposition of SBS *de novo* mutational signatures to the Catalogue of Somatic Mutations in Cancer (COSMIC) reference signatures was performed with SigProfilerAssignment (https://github.com/AlexandrovLab/SigProfilerAssignment) using the SBS-1536 *de novo* signatures with custom parameters. For SBS signatures this differed from the default parameters in two ways. Firstly, we increased the threshold where a signature is considered novel from 0.80 to 0.95, meaning that if the cosine similarity of the reconstructed signature compared to the *de novo* signature was less than 0.95, then the signature was considered novel. Secondly, we used the signature subgroups parameter to exclude signatures which were not likely to be present based on a combination of manual review of individual mutational spectra and prior knowledge of the biological mechanisms of the reference COSMIC signatures. The following groups of signatures were excluded; POL deficiency signatures (SBS10a, SBS10b, SBS10c, SBS10d, SBS28), homologous recombination (HR) deficiency signatures (SBS3), base excision repair (BER) deficiency signatures (SBS30, SBS36), iatrogenic signatures (SBS11, SBS25, SBS31, SBS32, SBS35, SBS86, SBS87, SBS90), ultraviolet (UV) signatures (SBS7a, SBS7b, SBS7c, SBS7d, SBS38), lymphoid signatures (SBS9, SBS85, SBS86), and artefact signatures (SBS27, SBS43, SBS45, SBS46, SBS47, SBS48, SBS49 , SBS50, SBS51, SBS52, SBS53, SBS54, SBS55, SBS56, SBS57, SBS58, SBS59, SBS60). These changes were necessary due to two well established issues with

mutational signature analysis, the presence of flat signatures and the large pool of reference SBS signatures.

Both previous studies and ours have shown that SBS5 and SBS40 are present in clear cell renal cell carcinomas (ccRCC).[2,3] These two mutational signatures are flat signatures, which in addition to being difficult to extract (as discussed above), are equally, if not more, difficult to accurately attribute. In addition to this, the large number of SBS-96 signatures in the current COSMIC reference set means that it is increasingly difficult to declare a mutational signature novel, as the majority of decompositions can achieve a cosine similarity of 0.80 using combinations of the existing reference signatures. Indeed, in our initial results all SBS *de novo* signatures could be decomposed using the default parameters, but in several cases the combination of signatures used was not plausible. For example, the decomposition result for signature SBS1536A (*aka,* SBS40b) using default parameters includes COSMIC reference signature SBS7c, which is associated with UV light exposure. This scenario is unlikely given that there is no plausible biological mechanism that could explain the presence of this signature in this cancer type. These problems could potentially be alleviated in the future by extending the contexts of the existing COSMIC reference signatures from SBS-96 to SBS-288 or SBS-1536, although such an analysis was outside the scope of current study.

For SBS signatures, increasing the cosine similarity threshold to 0.95, in combination with restricting the pool of signatures available for the decomposition as described above, results in five signatures (SBS_A, SBS_B, SBS_F, SBS_H, and SBS_I) remaining non-decomposed (**Fig.2, Supplementary Table 5**). For ID signatures, the decision was made to increase the threshold for a novel signature to 0.90, which results in a single signature (ID_C) remaining non-decomposed (**Extended Data Fig.4, Supplementary Table 5**). No alternation was deemed necessary for the decomposition of DBS78 *de novo* signatures, with one signature (DBS_D) remaining non-decomposed on default parameters (**Extended Data Fig.3, Supplementary Table 5**).

Justification for non-decomposed signatures

Although deviating from the default parameters during decomposition is an arbitrary decision, we consider this justified in the light of the additional evidence supporting the fact that these signatures genuinely represent distinct mutational processes.

Due to the nature of the flat signatures, including COSMIC reference signatures SBS5 and SBS40, it has been speculated that they do not represent a singular mutagenic process but instead multiple processes which are extremely difficult to separate due to their high level of correlation. Our results extracted 4 flat signatures. One of these, SBS_E is decomposed whereas SBS_A, SBS_B, and SBS_F remained non-decomposed. We provisionally have named these SBS40a, SBS40b, and SBS40c based on several pieces of evidence. Specifically, we did not extract a single mutational signature that directly matched SBS40 in any of the three extractions performed, whereas the combination of SBS40a, SBS40b, and SBS40c matches the COSMIC SBS40 with a cosine similarity of 0.96 (**Supplementary Fig. 7**). However, given the nature of COSMIC SBS40, cosine similarity is not the most robust piece of evidence given how readily flat signatures can be reconstituted. In addition, decomposing to SBS96 reference signatures results in the loss of differences found in the extended contexts which allowed the separation of the signature during the extraction process, and until such a reference set exists, comparing flat signatures to their reference set counterparts is suboptimal. Additional more convincing evidence comes from the fact that SBS40b (and to a lesser extent SBS40a) is associated with incidence, whereas SBS40c is not, in addition to the unique association of SBS40b with N, N, N-trimethyl-L-alanyl-L-proline betaine (TMAP). These findings imply that these are distinct mutational processes which have biologically meaningful implications in ccRCC. There are multiple reasons that could explain why SBS40 has not been split in previous analysis. Firstly, the majority of previous studies have utilised the SBS-96 context, yet the SBS-288 results in particular reveal substantial differences in the bias between genic and intergenic regions between SBS40a, SBS40b, and SBS40c. Secondly, our study has a much larger ccRCC cohort than was included in the study

which extracted the current COSMIC reference cohort. Thirdly, the fact that SBS40b alone associates strongly with incidence rate likely creates subtle differences in the ratio of the signatures to each other, which may have assisted in the extraction process.

For signatures SBS_I, ID_C, and DBS_D there is strong evidence to consider these new signatures which are associated with Aristolochic exposure. SBS_I was extracted in both SigProfilerExtractor 288 and 1536 formats, and while the mSigHdp version of the signature was not an exact match it did nonetheless agree that there was a second SBS signature in addition to the COSMIC signature 22. It is also possible to identify mutational spectra from individual tumours which are dominant in both signatures (**Supplementary Fig. 8**). All four signatures correlate strongly with each other, and show enrichment in the same countries (Romania, Serbia and Thailand, **Extended Data Fig.5-6**). Finally, SBS_H is justified given the presence of a hypermutator whose mutational spectra is a close match to the signature (**Supplementary Fig. 4**). The cancer of this individual has a mutational burden exceeding the mutational burdens of cancers with the strongest Aristolochic exposure; however, with only a single case, it is impossible to determine whether this is due to an environmental exposure or due to a currently unknown defect in DNA repair.

Presence of tobacco-associated signature SBS4

Whilst tobacco has been shown to be a risk factor for RCC, SBS4 has not been previously identified in RCC.[2,4] In this study SBS4 was identified as a component of the *de novo* signature SBS_C, which is decomposed into COSMIC signatures SBS4 and SBS40. To provide confidence in the presence of SBS4 the following steps were performed. Firstly, the signature was re-decomposed using a custom COSMIC reference signature list where SBS40 is replaced with SBS40a, SBS40b, and SBS40c. This shows that the *de novo* signature is composed solely of SBS4 (34.52%) and SBS40a (65.48%). Whilst it is not possible to completely remove the SBS40a component, subtracting the SBS40a contexts at the above ratio should leave a signature which is a closer match to the COSMIC reference SBS4, and

indeed, the resulting adjusted signature has a cosine similarity of 0.90 compared to 0.80 in the original *de novo* signature (**Supplementary Fig. 9**). The adjusted signature notably lacks the T>A peaks present in the reference SBS4, whilst comparing just the C>A compartment increases the cosine similarity of the adjusted signature even further to 0.95. In a previous study of environmental exposures, which included many compounds found in tobacco smoke, the compounds dibenzo[*a,l*]pyrene diol-epoxide (DBPDE) and dibenzo[*a,l*]pyrene (DBP) generated a profile which strongly resembles the T>A peak observed in SBS4 (**Supplementary Fig. 9**).[5] We can speculate that the absence of these peaks in the adjusted signature indicates that not all mutagenic components of tobacco smoke are present in the kidney, however this would require additional study to confirm. For the purposes of this study, the adjusted signature provides sufficient evidence of an SBS4-like signature in ccRCC.

Attribution of single base substitution signatures

For *de novo* SBS signatures, the decision was made to exclude SBS_H. This signature was driven by a single hypermutator, without which the signature is not extracted. However, when included in the attribution panel, this signature present in 208 /962 cases (22%) likely due to overlap in contexts with multiple other signatures. Therefore, SBS_H was removed from the final *de novo* attribution panel. For COSMIC reference SBS signatures, attributions were performed on the subset of COSMIC reference signatures which are present in the dataset (as determined by SigProfilerAssignment following decomposition of the *de novo* signatures), in addition to any non-decomposed signatures. Two changes were made to this for the final attributions. Specifically, SBS40 was removed from the panel of signatures. SBS40 was found in several decompositions, likely where there is a low-level background of mutations in the *de novo* signatures. However, it does not make sense to include it the final panel given the presence of SBS40a, SBS40b and SBS40c. Additionally, SBS_H was removed from the final COSMIC attribution panel for the same reasoning as for *de novo* signatures.

Attribution of SBS40 components in a pan-cancer cohort

In order to investigate the patterns of attribution of SBS40a, SBS40b, and SBS40c, SigProfilerAssignment was used to attribute an altered COSMIC reference set where SBS40 is replaced with SBS40a, SBS40b, and SBS40c on a pan-cancer cohort dataset.[3] This showed that SBS40a was found in the majority of tumour types, whilst SBS40b and SBS40c were only seen consistently in clear cell RCC (**Supplementary Fig. 10**). Notably, the chromophobe RCC dataset did not show attribution to either SBS40b or SBS40c, suggesting that these signatures are likely further restricted to certain niches within the kidney. These results provide additional support for SBS40a, SBS40b, and SBS40c representing distinct mutational processes.

Attribution of SBS12 in liver cancers

The COSMIC reference signature SBS12 was originally extracted from liver cancers.[2,3] The liver cancers used in this extraction consisted of three cohorts, one of which (LINC-JP) is from Japan. In order to determine whether SBS12 enrichment is also present in Japanese liver cancers, SigProfilerAssignment was used to attribute the COSMIC v3.3 reference signatures. This shows that SBS12 is enriched in the LINC-JP cohort compared to the LICA-FR (France) and LIHC-US (US) cohorts (**Supplementary Fig. 11**).

SUPPLEMENTARY REFERENCES

1. Liu, M., Wu, Y., Jiang, N., Boot, A. & Rozen, S. G. mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery. *NAR Genom Bioinform* **5**, (2023).
2. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
3. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
4. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science (1979)* **354**, 618–622 (2016).
5. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Note Table 1: Comparison of single base substitution signatures extracted by SigProfilerExtractor and mSigHdp.

Supplementary Fig. 1: Single base substitution signatures extracted by SigProfilerExtractor.
All single base substitution (SBS) *de novo* signatures extracted in SBS-288 (11 signatures) and SBS-1536 (13 signatures) format, shown side by side for comparison. Equivalent signatures where not extracted in SBS-288 format for SBS1536C and SBS1536E. For clarity, the signatures context is retained in the signature names in this figure.

Supplementary Fig. 2: Doublet base substitution signatures extracted by SigProfilerExtractor.
Four doublet base substitution (DBS) *de novo* signatures extracted by SigProfilerExtractor.

Supplementary Fig. 3: Small insertion and deletion signatures extracted by SigProfilerExtractor.
Seven small insertion and deletion (ID) *de novo* signatures extracted by SigProfilerExtractor

Supplementary Fig. 4: Single base substitution mutational signature driven by a hypermutated kidney cancer.
(a) A single base substitution signature extracted in SBS-1536 format (SBS_H) and **(b)** the mutational spectra of a clear cell renal cell carcinomas (ccRCC) patient which corresponds to the extracted signature. The mutation burden in this patient was the highest observed in the cohort.

Supplementary Fig. 5: Single base substitution mutational signatures extracted by mSigHdp.
Eleven single bases substitution (SBS) *de novo* signatures extracted by mSigHdp.

Supplementary Fig. 6: Small insertion and deletion mutational signatures extracted by mSigHdp.
Six small insertion and deletion (ID) *de novo* signatures extracted by mSigHdp.

Supplementary Fig. 7: Reconstruction of COSMIC reference signature SBS40.
The combination (SBS_ABF) of *de novo* signatures SBS_A, SBS_B and SBS_C at equal ratios **(a)** can reconstruct the profile of COSMIC signature SBS40 **(b)** with a cosine similarity of 0.96.

Supplementary Fig. 8: Aristolochic acid mutational signatures in kidney cancers.
Examples of individual RCC mutational spectra which support the existence of both SBS22a **(a)** and SBS22b **(b)**.

Supplementary Fig. 9: Presence of tobacco-associated signature SBS4 in in kidney cancers.
SBS4 was identified as a component of SBS_C **(a)** which also contains SBS40a. Subtracting the SBS40a component results in SBS_C_adjusted **(b)** which has a higher overall cosine similarity (CS) to COSMIC reference signature SBS4 **(c)**. The previously determined mutational signatures of the compounds dibenzo[*a,l*]pyrene (DBP) **(d)** and dibenzo[*a,l*]pyrene diol-epoxide (DBPDE) **(e)** generate peaks which correspond to the T>A peaks observed in SBS4, and the absence of these compounds in kidney may explain the remaining difference of SBS_C_adjusted compared to SBS4.

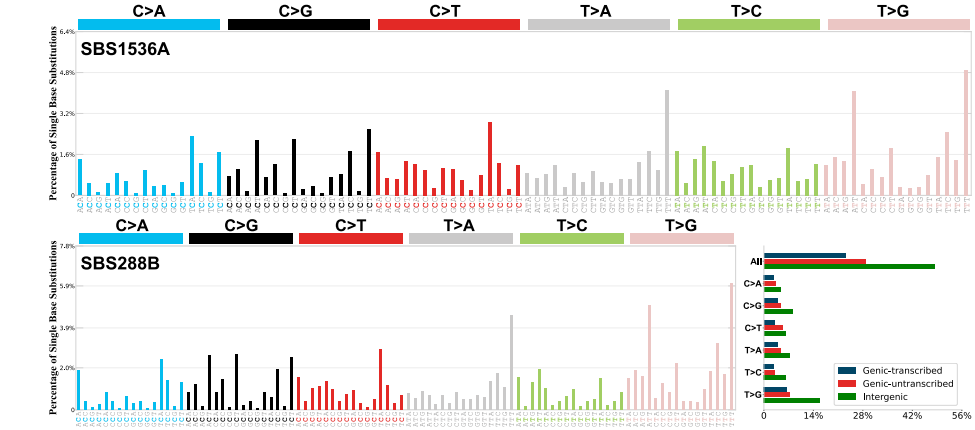
Supplementary Fig. 10: Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-cancer cohort.
Attribution of signatures SBS40a, SBS40b, and SBS40c in a pan-cancer cohort, showing a widespread distribution for SBS40a whilst SBS40b and SBS40c are only seen consistently in clear cell renal cell carcinomas (ccRCC).

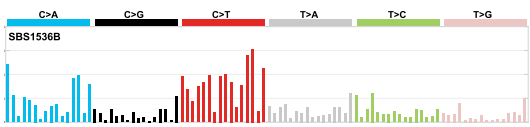
Supplementary Fig. 11: Attribution of signature SBS12 in liver cancers
Attribution of SBS12 in liver cancers, showing enrichment of COMSIC reference signature SBS12 in the LIRI-JP cohort (Japan) compared to those in LIRI-US (USA) and LIRI-FR (France) cohorts.

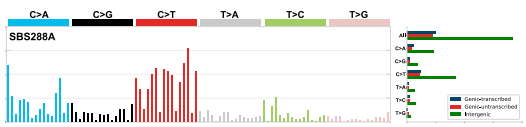
SBS1536 Signature	SBS288 Signature	mSigHdp Signature	SBS1536 vs SBS288 Cosine Similarity	SBS1536 vs mSigHdp Cosine Similarity	SBS288 vs mSigHdp Cosine Similarity
SBS1536A	SBS288B	hdp2	0.98	0.95	0.89
SBS1536B	SBS288A	hdp1	0.97	0.95	0.89
SBS1536C	-	-	-	-	-
SBS1536D	SBS288F	hdp3	1	0.99	0.98
SBS1536E	-	-	-	-	-
SBS1536F	SBS288C	hdp4	0.99	0.95	0.9
SBS1536G	SBS288E	hdp9	0.99	0.94	0.95
SBS1536H	SBS288D	hdp10	0.98	0.97	0.93
SBS1536I	SBS288J	hdp 11	1	0.75	0.77
SBS1536J	SBS288H	hdp6	0.99	0.92	0.92
SBS1536K	SBS288K	hdp7	0.99	0.97	0.99
SBS1536L	SBS288I	hdp5	0.99	0.99	0.98
SBS1536M	SBS288G	hdp8	0.98	0.96	0.95

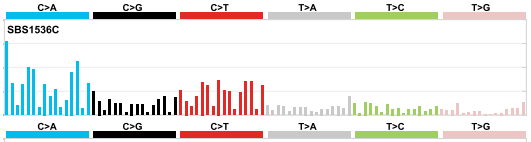
Supplementary Note Table 1: Comparison of single base substitution signatures extracted by SigProfilerExtractor and mSigHdp

Supplementary Fig.1: Single base substitution mutational signatures extracted by SigProfilerExtractor

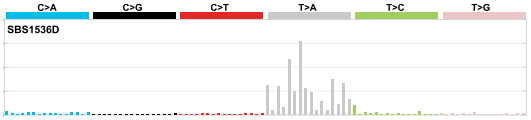


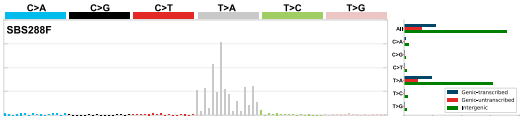


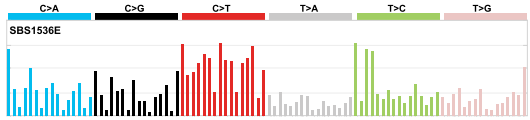




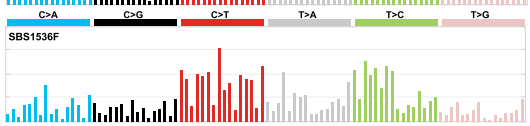
no direct counterpart

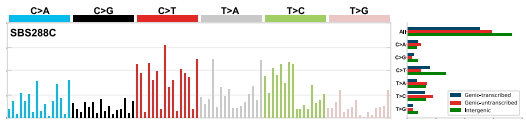


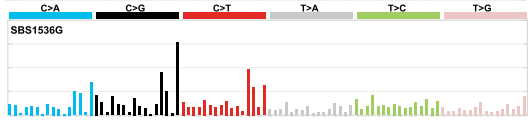


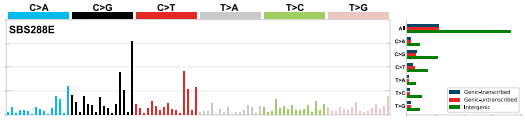


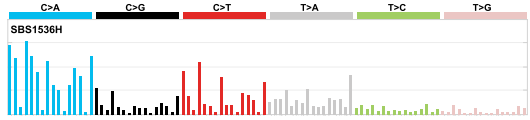
no direct counterpart

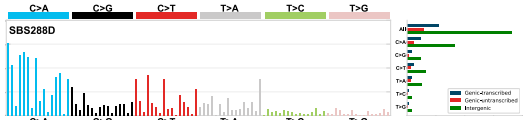


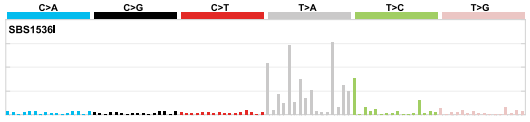


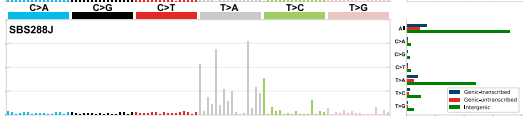


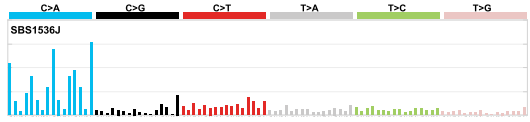


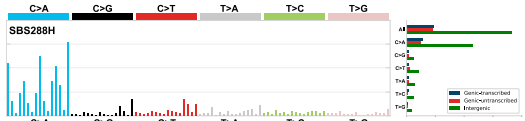


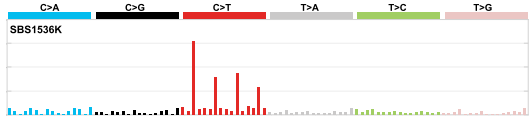


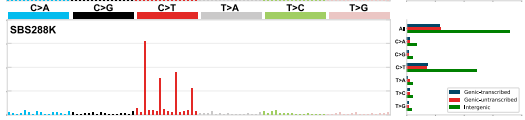


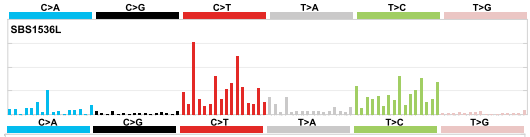


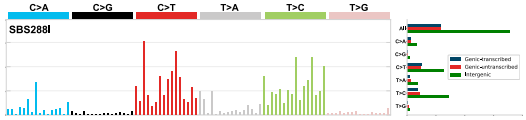


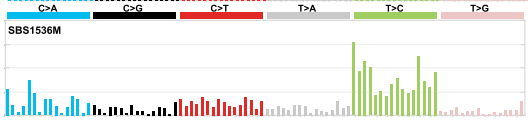


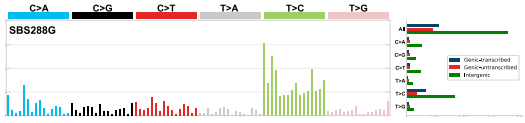






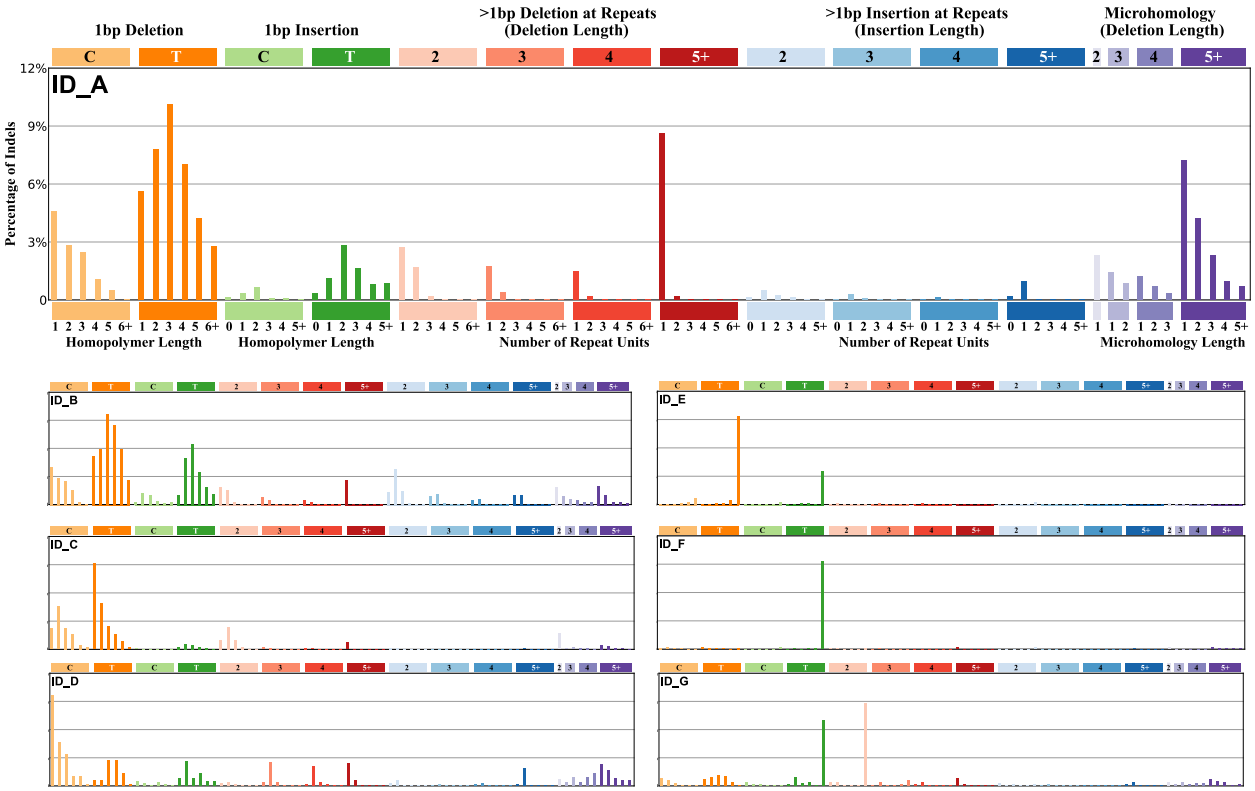






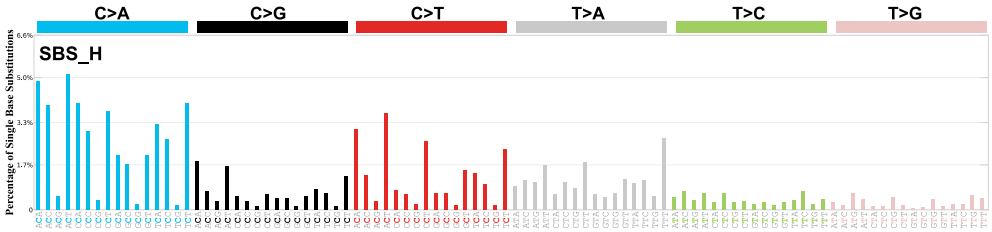
Supplementary Fig.2: Doublet base substitution mutational signatures extracted by SigProfilerExtractor

Supplementary Fig.3: Small insertion and deletion mutational signatures extracted by SigProfilerExtractor

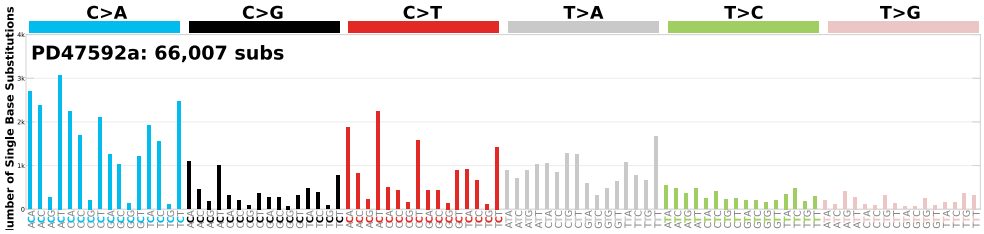


Supplementary Fig.4: Single base substitution mutational signature driven by a hypermutated kidney cancer

A

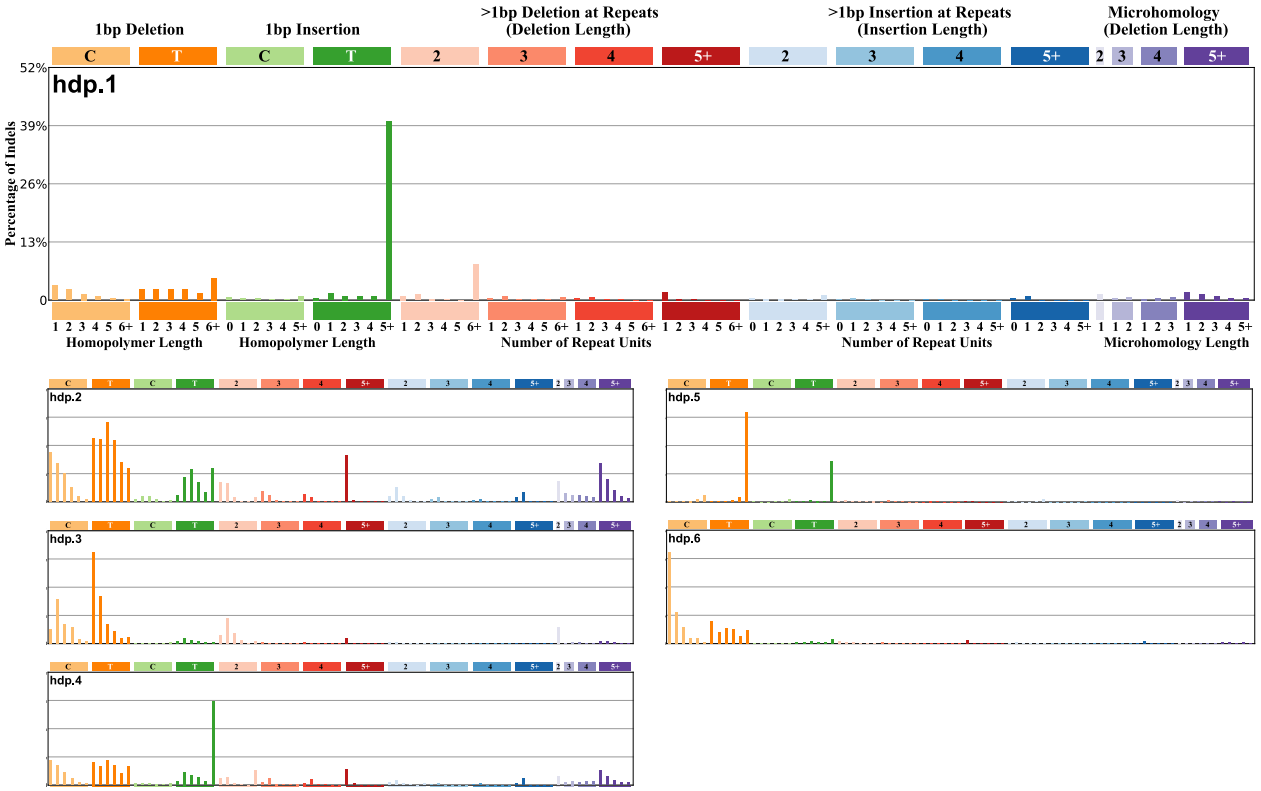


B



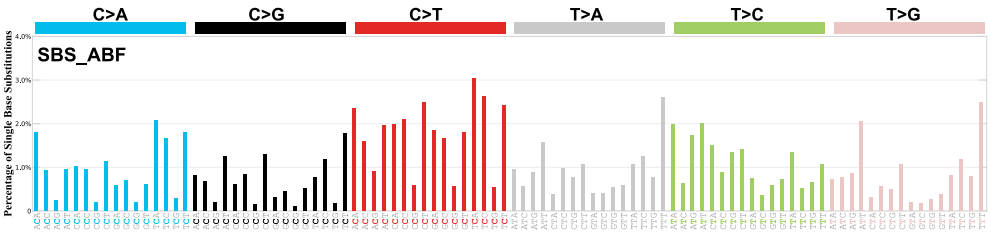
Supplementary Fig.5: Single base substitution mutational signatures extracted by mSigHdp

Supplementary Fig.6: Small insertion and deletion mutational signatures extracted by mSigHdp

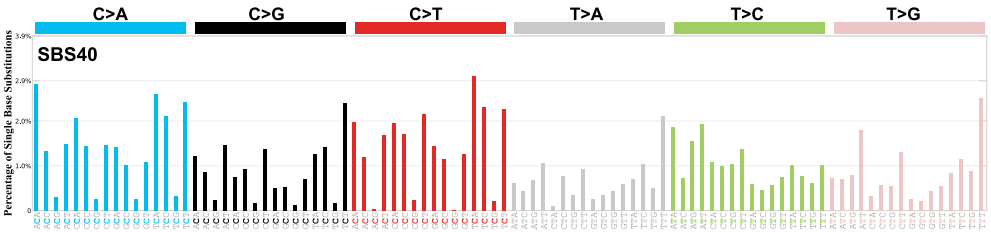


Supplementary Fig.7: Reconstruction of COSMIC reference signature SBS40

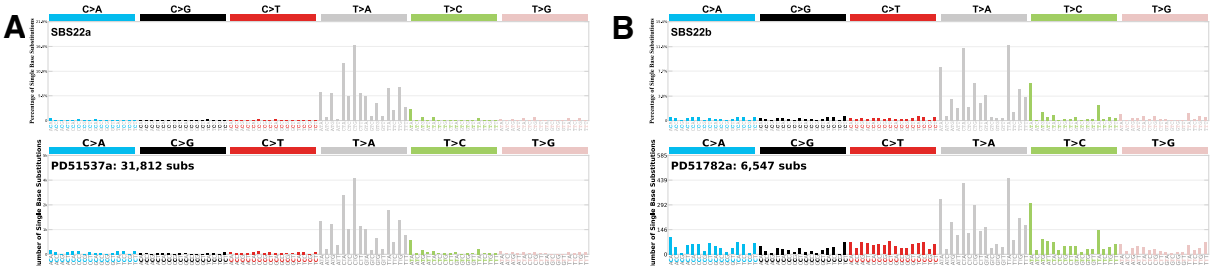
A



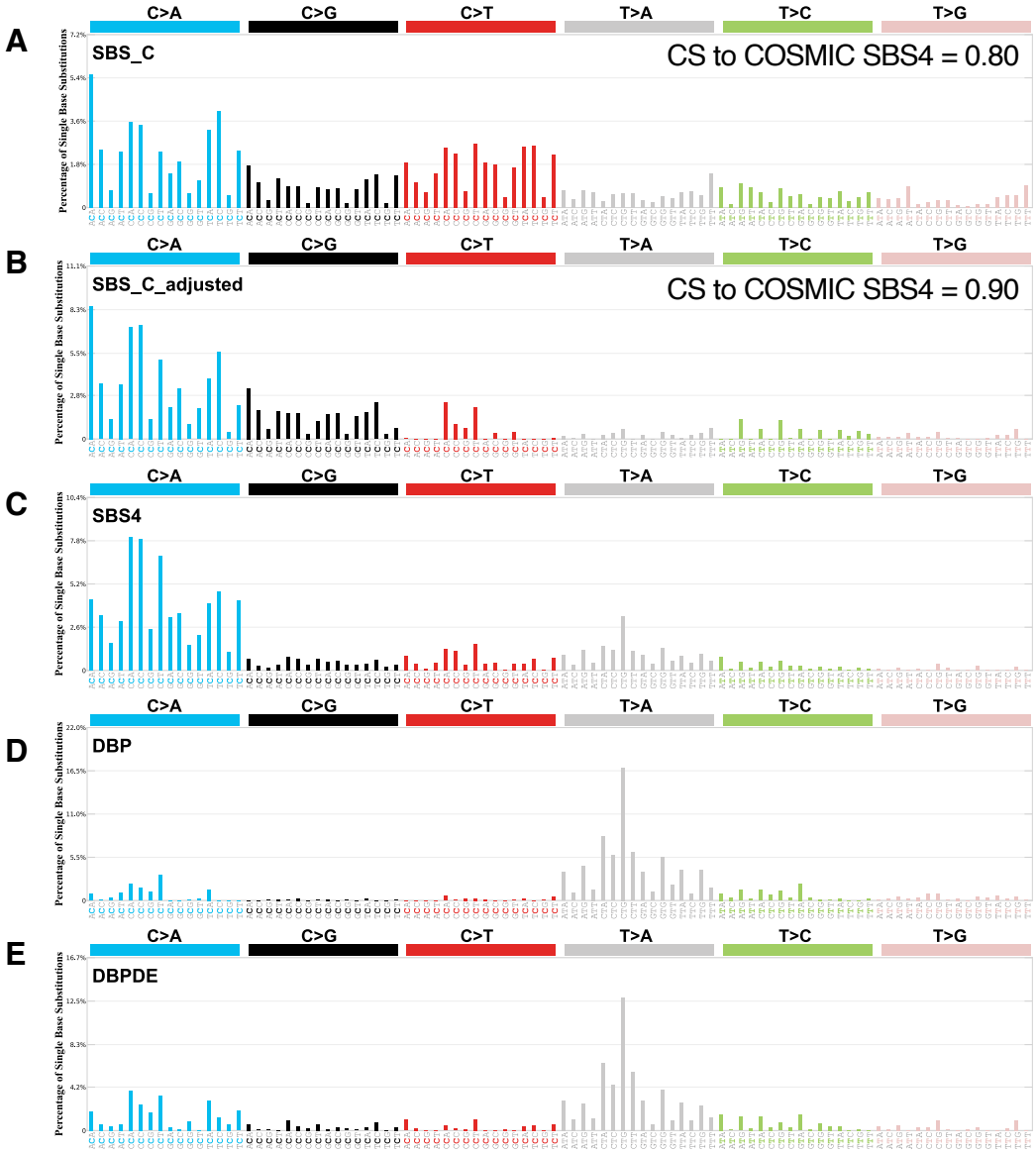
B



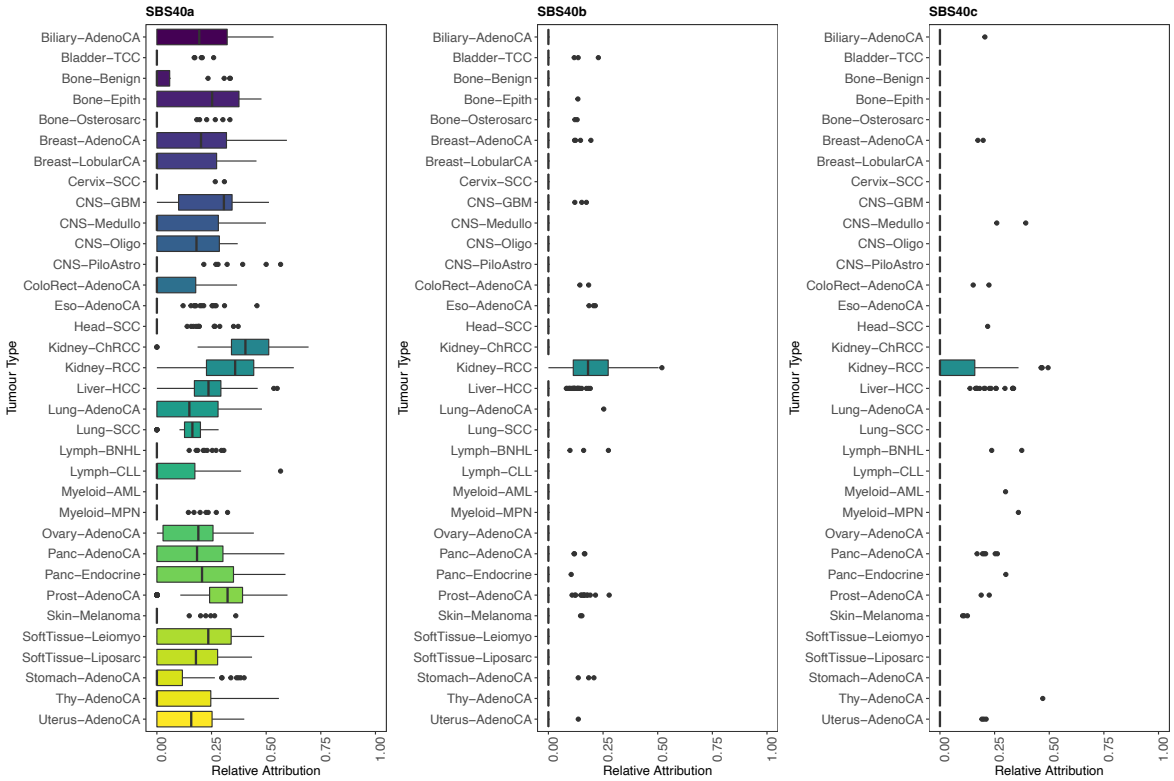
Supplementary Fig.8: Aristolochic acid mutational signatures in kidney cancers



Supplementary Fig.9: Presence of tobacco-associated signature SBS4 in in kidney cancers



Supplementary Fig.10: Attribution of mutational signatures SBS40a, SBS40b and SBS40c in a pan-cancer cohort



Supplementary Fig.11: Attribution of signature SBS12 in liver cancers