Mutational signature analysis of Asian OSCCs reveals novel mutational signature with exceptional sequence context specificity

Arnoud Boot[1,2], Alvin W.T. Ng[2,3], Fui Teen Chong[4], Daniel S.W. Tan[4], N. Gopalakrishna Iyer[1,4], Steven G. Rozen[1,2,3]*

[1] Cancer and Stem Cell Biology, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[2] Centre for Computational Biology, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[3] NUS Graduate School for Integrative Sciences and Engineering, 28 Medical Drive, 117456, Singapore

[4] Cancer Therapeutics Research Laboratory, Division of Medical Science, National Cancer Centre Singapore, 169610, Singapore

* Corresponding author: SGR: steve.rozen@duke-nus.edu.sg

Abbreviations:

bp	base pair
indels	insertions and deletions
SBS	single base substitutions
DNS	dinucleotide substitutions
HNSCC	head and neck squamous cell carcinoma
OSCC	oral squamous cell carcinoma
PCAWG	Pan-Cancer Analysis of Whole Genomes
TC-NER	transcription coupled nucleotide excision repair

Keywords:

Oral squamous cell carcinoma, mutational signature, sequence context specificity, duocarmycin

Novelty and impact:

Here, we describe a novel mutational signature with exceptional enrichment for adenines in the 4 base pairs 5' of mutation sites. We first observed this signature in an oral squamous cell cancer in which it dominated, and then detected it in 23 of 23,829 tumours that had publicly available data. Identification of rare mutational signatures can lead to improved understanding of carcinogenesis and carcinogenic compounds, thereby leading to improved prevention strategies.

Abstract

Study of mutational signatures in cancers can reveal the history of mutagenic processes that cells were exposed to prior and during tumourigenesis. Although recent work has identified 65 distinct single nucleotide substitution signatures, we expect that as-yet-undiscovered mutational processes will be able to shed further light on mutagenesis leading to carcinogenesis. For this reason we analyzed mutational signatures in whole-exome sequencing data of 36 Asian oral squamous cell carcinomas.

The mutational spectra of most of the carcinomas could be explained by known mutational signatures, but one sample showed a novel mutational signature. Whole-genome sequencing revealed that this mutational signature was characterized by a preponderance of thymine mutations, strong transcriptional strand bias, and a striking enrichment for adenines in the 4 base pairs 5' of mutation sites. The enrichment was strongest 3 base pairs 5' of mutated thymines, where 93.5% of bases were adenines. We also observed deletions of single thymines inside and outside of thymine repeats with similar 5' enrichment for adenines. Examination of publicly available whole-genome and whole-exome sequencing data of 23,829 tumours revealed this signature in 23 tumours from sites including the bile-duct, bladder, pancreas and rectum.

Although the aetiology of this mutational signature remains unknown, the transcriptional strand bias suggests adduct formation on adenines. We speculate that the signature may have been caused by duocarmycins, naturally occurring large DNA-crosslinking molecules, which were also used in the past in clinical trials. We call this novel mutational signature SBS_A^nT.

Introduction

Mutagenesis is one of the major causes of cancer. A thorough mapping of mutational signatures promises to illuminate the mechanisms of carcinogenesis and help identify carcinogenic mutagenic compounds or processes. In recent years, the field of mutational-signature analysis has made huge strides in identifying distinct mutational processes. Currently, 65 distinct single base substitution signatures have been described.[1] Most of these stem from defects in DNA repair and replication, endogenous mutagenic processes, or exposure to mutagenic compounds such as benzo[a]pyrene or aristolochic acid. However, for 20 mutational signatures the aetiology remains unknown.[1]

Although most common mutational processes are represented by known mutational signatures, we expect that there are additional mutational processes that contribute to only small subsets of tumours. An example of such a rare signature is SBS42, due to occupational exposure to haloalkanes.[1, 2] This signature was not discovered in the original COSMIC signatures,[3] but was only discovered in cholangiocarcinomas from patients who worked at a printing company. SBS42 was extremely rare in other cancer types.[1] This example suggests that there are more rare mutational processes that are due to rare occupational exposures, dietary exposures, or genetic variants affecting DNA repair or replication mechanisms. Such mutational processes will be challenging to find, but they will point to cancers that could be prevented if the responsible mutagens can be identified and exposure to them avoided.

Head and neck squamous cell carcinoma (HNSCC) is the 6th most common cancer worldwide, with more than 680,000 new cases every year.[4] With 300,000 new cases per year, oral squamous cell carcinoma (OSCC) is the largest subtype.[4] In OSCCs, 9 different mutational signatures have been detected, but >92% of mutations are due to mutational signatures associated with aging (clock-like signatures SBS1 and SBS5), APOBEC cytidine deaminases (SBS2 and SBS13), and chewing tobacco (SBS29).[1] To search for possible rare mutational processes we analyzed whole-exome sequencing data of 36 Asian OSCCs.

Materials and Methods

Samples

De-identified tissue samples and matching whole-blood were collected from OSCC patients operated on between 2012 and 2016 at the National Cancer Centre Singapore. In accordance with the Helsinki Declaration of 1975, written consent for use of clinical material and clinico-pathologic data for research was obtained at the time of surgery. This study was

approved by the SingHealth Centralized Institutional Review Board (CIRB 2007/438/B). Sample information is included in Supporting Information Table S1.

Whole-exome and whole-genome sequencing

Whole-exome sequencing was performed at Novogene. (Beijing, China) on a HiSeq X Ten instrument with 150bp paired-end reads. Whole-genome sequencing was performed at BGI (Hong Kong) on the BGIseq500 platform, generating 100bp paired-end data.

Alignment and variant calling

Sequencing reads were trimmed by Trimmomatic[5]. Alignment and variant calling and filtering was performed as described previously.[6]

Validation of SBSs by Sanger sequencing

We selected 96 variants with >15% variant allele frequency, and excluded variants immediately adjacent to a homopolymer of ≥9 bp. PCR product purification and Sanger sequencing was performed at GENEWIZ© (Suzhou, China).

Signature assignment

We assigned mutational signatures to the mutational spectra of the 30 OSCCs with ≥10 mutations using SigProfiler and the SigProfiler reference mutational signatures.[1] As OSCC is a subset of HNSCC, all mutational signatures that were identified in HNSCCs and OSCCs in the International Cancer Genome Consortium's Pan Cancer Analysis Working Group (PCAWG) analysis were included for reconstruction.[1] The PCAWG mutational signatures were adjusted to exome trinucleotide frequency where required.

Gene expression data

Single cell gene expression data for OSCC was downloaded from NCBI GSE103322.[7] We took the median gene expression for all tumour cells as the representative expression level of OSCCs.

Detection of SBS_A^{n}T in publicly available sequencing data

19,184 whole-exomes and 4,645 whole-genomes data previously compiled for the PCAWG mutational signature analysis were screened for presence of SBS_AnT, including 2,780 PCAWG whole-genomes, and 9,493 exomes from the TCGA consortium.[1, 8] We looked for tumours with ≥50 (exomes) or ≥500 (genomes) thymine mutations, and with enrichment for mutations with the 5' sequence context characteristic of SBS_A^{n}T (Supporting Information Data S1).

We then used mSigAct to test for SBS_A^{n}T amongst the candidate tumours identified in the previous step (Supporting Information Data S2).[6, 9] mSigAct provides a *p* value for the null hypothesis that a signature is not needed to explain an observed spectrum compared to the alternative hypothesis that signature is needed.

Results

Mutational signature analysis

We analyzed whole-exome sequencing data from 36 OSCCs treated in Singapore, including 18 previously published OSCCs.[10] These tumours had significantly fewer somatic single base substitution (SBSs) than previously published OSCCs and HNSCCs (median 1.02 versus 1.66 and 2.44 mutations per megabase, $p = 4.11\times10^{-5}$ and 4.85×10^{-10} respectively, Wilcoxon rank sum tests).[1] We excluded 6 tumours from further analysis because they had < 10 SBSs. The mutational spectra of the remaining 30 tumours are shown in Supporting Information Figure S1.

We computationally reconstructed the mutational spectra of the 30 tumours using the mutational signatures previously observed in HNSCCs and OSCCs (Supporting Information Figure S2A).[1] We observed evidence of APOBEC mutagenesis in 17 tumours. Other observations included evidence of mutagenesis associated with MUTYH and NTHL1 deficiency. Strikingly, two samples with high mutation loads, 62074759 and TC1, were poorly reconstructed (Supporting Information Figure S2B). These mutational spectra showed very distinctive mutation patterns. This led us to hypothesize that each was caused predominantly by a single, novel, mutational process, which in the case of TC1 appeared to be combined with APOBEC mutagenesis (Supporting Information Figure S1)[1]. Both spectra showed T>A and T>C peaks with strong transcriptional strand bias, but were clearly distinct.

Characterization of the mutational spectrum in 62074759

During routine visual inspection of the read alignments supporting the somatic variants in 62074759, we noticed that 51 out of the 84 T>C mutations were directly preceded by at least 3 adenines (3 adenines directly 5' of the T>C mutation). In addition, most of the TTT>TNT mutations were located within TTTT homopolymers. Because of the high risk of sequencing errors in and near homopolymers, we validated 96 somatic SBSs detected in 62074759, all of which were confirmed.

We next sequenced the whole-genome of 62074759, identifying 34,905 somatic SBSs and 4,037 small insertions and deletions (indels). The whole-genome SBS mutation spectrum confirmed the spectrum observed in the exome (Fig. 1A, Supporting Information Figure S1). The spectrum was dominated by AT>AA and AT>AC mutations with a main peak at ATT>ACT, and by TTT>TAT, TTT>TCT and TTT>TGT mutations. The genome data showed the same striking enrichment for adenines 5' of T>C mutations as the exome data. Among **all** SBSs, 79.5% had an adenine 3bp 5' of the mutation sites and 65.3% had an adenine 4bp 5' of the mutation (Fig. 1B). The subset of SBSs from thymine showed stronger enrichment; 93.5% and 75.2% had adenines 3bp and 4bp 5' of the mutation, respectively. Among the even more restricted subset of ATT>ACT SBSs, 98.2% had an adenine 3bp 5' of the mutation. We also observed strong enrichment for AAA 5' of SBSs (Fig. 1C). No enrichment of adenine 5' of mutated cytosines was observed (Supporting Information Figure S3). The variant allele frequency of SBSs was consistent with prolonged exposure (Supporting Information Figure S4).

Mutational processes associated with large adducts are known to generate more mutations on the non-transcribed strands of genes than on the transcribed strands, due to transcription-coupled nucleotide excision repair (TC-NER) of the adducts on transcribed strands.[1] To investigate whether this novel signature might have been caused by large adducts, we examined its transcriptional strand bias. We observed very strong strand bias on all classes of thymine mutations. The enrichment of mutations when thymine is on the transcribed strand (and adenine is on the non-transcribed strand) is indicative of adduct formation on adenines. Also consistent with the activity of TC-NER, transcriptional strand bias of T>A, T>C and T>G mutations correlated strongly with transcriptional activity (Fig. 1D, $p = 9.50\times10^{-41}$, 6.33×10^{-91} and 5.69×10^{-33} respectively, Chi-squared tests). Furthermore, TC-NER proficiency decreases with increasing distance to the transcription start site.[6, 11, 12] Consistent with this, T>A, T>C and T>G mutations all showed decreased transcriptional strand bias towards the 3' end of transcripts (Fig. 1E, $p = 4.32\times10^{-32}$, 1.26×10^{-78} and 6.64×10^{-24} respectively, logistic regression). None of the cytosine mutation classes showed transcriptional strand bias. This, plus the absence of enrichment for adenines 5' of cytosine

mutations, suggests that the cytosine mutations in this sample were not caused by the same mutational process as the thymine mutations.

Insertions, deletions and dinucleotide substitutions

The vast majority of indels were deletions (98.6%), mainly of single thymines (Fig. 2A). The indel spectrum did not resemble any of the previously published indel signatures.[1] Like the SBSs, deletions of thymines in thymine mono-and dinucleotides showed strong enrichment for three preceding adenines (Fig. 2B+C). Thymine deletions in thymine tri- to octonucleotides, had very strong enrichment for a single adenines immediately 5' of the thymine repeat, but enrichment for adenines further 5' decreased rapidly for longer repeats (Supporting Information Figure S5). This suggests adduct formation opposite the thymine at the 3rd or 4th thymine of longer thymine repeats. For thymine deletions outside of thymine repeats, we observed strand bias consistent with adenine adducts (p = 0.01, binomial test, Supporting Information Figure S6). Contrastingly, thymine deletions in thymine tetranucleotides showed transcriptional strand bias in the opposite direction (p = 0.01, binomial test), and thymine deletions in thymine homopolymers of other lengths lacked transcriptional strand bias.

We detected 171 dinucleotide substitutions (DNSs) in tumour 62074759, most of which were CC>NN and TC>NN substitutions (Supporting Information Figure S7). As the predominant mutational process in 62074759 caused almost exclusively thymine mutations, it is unlikely that these DNSs were caused by the same mutational process. The DNS spectrum best resembles the DNS spectrum observed in cisplatin exposed cell lines (cosine similarity of 0.859).[6] Concordantly, patient 62074759's clinical records showed that prior to the development of the tumor that we sequenced, which was a recurrence, the initial tumour had been treated preoperatively with several chemotherapies, including cisplatin.

SBS_AnT in publicly available sequencing data

To investigate whether the mutational signature from 62074759 (hereafter referred to as SBS_A^{n}T) was also present in other tumours, we investigated 4,645 tumour genomes and 19,184 tumour exomes compiled for the PCAWG Mutational Signatures Working Group.[1] We identified 39 whole-exome and 16 whole-genome samples with possible presence of SBS_A^{n}T based on enrichment for adenines 3 and 4bp 5' of thymine mutations (Supporting Information Data S1). These included tumours of the bladder, colon, rectum, and prostate. Supporting Information Figure S8 shows their mutational spectra.

To further narrow down the selection of tumours showing SBS_A^{n}T we used the mSigAct signature presence test to assess whether SBS_A^{n}T was needed to explain the observed spectrum (Supporting Information Data S2).[9] Using the signature assignments previously reported for these tumours[1], we compared reconstruction of the mutational spectra with and without SBS_A^{n}T. We identified 23 tumours which were reconstructed significantly better when including SBS_A^{n}T (Fig. 3). Of these, indel information was only available for the PCAWG whole-genomes. Five out of these had thymine deletions with the expected sequence contexts (Supporting Information Figure S9+S10).

Characterization of the mutational signature in TC1

Whole-genome sequencing of TC1 identified 5,402 SBSs and 67 indels. Besides APOBEC-associated mutations, we observed prominent TG>AG peaks and a strong GTG>GCG peak, all with strong transcriptional strand bias (Supporting Information Figure S11). Unlike SBS_A^{n}T, for which we used the 5' enrichment for adenines to identify likely exposed tumours, for the signature in TC1, there was only the proportions of SBSs in trinucleotide context. Therefore we did a screen for cosine similarity between the thymine mutations, as well as the T>A mutations specifically for all 23,829 tumours. We found no tumours in which presence of the TC1 mutational signature was visible in the mutation spectrum (Supporting Information Figure S12).

Discussion

We identified a novel mutational signature, SBS_A^{n}T, which is characterized by T>A, T>C and T>G mutations with an exceptional preference for preceding adenines. We also detected SBS_A^{n}T in publicly available whole-exome and whole-genome mutation data of tumors from amongst others the digestive tract (colon, rectum, oesophagus, bile duct and pancreas) and the urinary tract (bladder and prostate). Thus, this signature is not tissue specific. We also detected evidence of another novel mutational signature, but found no evidence of this signature in any other tumour. Initially, we hypothesized that there were still rare mutational processes to be discovered. Our discovery of SBS_A^{n}T and the signature in TC1 bore this out. It appears that these signatures are indeed very rare, as we found only 23 additional tumours with SBS_A^{n}T and no additional tumours with the TC1 signature.

Patient 62074759 had an unusual treatment history. Although the usual first-line treatment for OSCC of the tongue is surgery, prior to becoming a patient at our institution, 62074759 was first treated with gemcitabine, cisplatin, 5-fluorouracil and radiotherapy. None

of these is likely to cause SBS_A^{n}T. Reported cisplatin and radiotherapy mutational signatures do not resemble SBS_A^{n}T.[6, 13] Gemcitabine acts as a cytidine analogue, and is therefore unlikely to cause thymine mutations. Lastly, 5-fluorouracil is primarily a thymidylate synthase inhibitor, causing a depletion in the cellular deoxythymidine monophosphate. This could possibly generate a mutational signature, but we would not expect transcriptional strand bias or such a striking sequence context specificity.

SBS_A^{n}T shows strong transcriptional strand bias, which is commonly observed for mutational processes associated with bulky adducts.[6, 9, 12] The depletion of thymine mutations on untranscribed strand suggests that the mutational process causing SBS_A^{n}T involves formation of a bulky adduct on adenine. Fig. 2D shows a proposed model for adduct formation leading to SBS_A^{n}T. The model assumes 2 independent adducts are formed, either inside adenine-homopolymers, or directly adjacent to thymine-homopolymers. The sequences for the adducts in the model are the reverse complement of each other, and we cannot exclude the possibility of an interstrand crosslink. However, if this were the case, we would expect to also observe multiple pairs of SBSs separated by 2 unaffected bases, which we did not.

Focussing on the characteristics of SBS_A^{n}T we did a literature search for compounds that could cause SBS_A^{n}T. A class of naturally occurring minor groove binding DNA alkylating agents called duocarmycins stood out. Several species of *Streptomyces* are known to produce duocarmycins.[14, 15] In the late 1990's and early 2000's, several synthetic duocarmycin analogues reached clinical trials that were discontinued due to toxicity.[16-18] Recently, antibody-drug conjugates with duocarmycins have revitalised research into their possible clinical uses.[19] SYD985, an trastuzumab-duocarmycin conjugate,[20] has recently entered phase III clinical trials for the treatment of metastatic breast cancer (clinicaltrials.gov/ct2/show/NCT03262935). Duocarmycins bind specifically to adenines in A/T-rich regions, which matches SBS_A^{n}T's sequence context.[21-23] Conceivably, patient 62074759 could have been exposed through contact with *Streptomyces,* as several species are known human symbionts.[24] However, it is unclear whether these species produce duocarmycins. Alternatively, we speculate that the patient, while residing outside of Singapore prior to treatment at our institution, may have been exposed to one of the duocarmycin analogues previously used in clinical trials.

Acknowledgements

The results here are partly based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/) and data assembled by the International Cancer Genome

Consortium Pan Cancer Analysis Working Groups. We thank Willie Yu for technical assistance.

Data availability

FASTQ files for all sequencing are at the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/home) under accession <data being uploaded now>.

Figure legends

Fig. 1: **In depth characterization of SBS_A^{n}T in the whole-genome data from tumour 62074759. (A)** SBS profile. **(B)** Sequence context preferences of all SBSs, revealing strong preference for an adenine 3bp 3' of mutated thymines. **(C)** Per-mutation view of sequence context preferences of mutations from thymines; each row represents one mutation with bases indicated by colour as in panel B. **(D)** Transcriptional strand bias as a function of gene expression level and distance to the transcription start site **(E)**.

Fig. 2: **In depth characterisation of the A^{n}T indel signature in the whole genome data from tumour 62074759. (A)** Spectrum of indels in the classification proposed by the PCAWG consortium.[1] The indel spectrum is dominated by deletions of single thymines (orange). The numbers at the bottom indicate the lengths of the repeats the deletions occured in. 1 denotes deletions of thymines flanked by non-thymimes, 2 denotes deletions of thymines in thymine dinucleotides; (TT>T) and so on. **(B)** Sequence contexts of deletions of a single thymine, not in a T-repeat (top) and in a thymine dinucleotide (bottom). Analogous plots for deletions of single thymines in homopolymers of 3 to 8 thymines are shown in Supporting Information Figure S5. **(C)** Per-mutation view of sequence contexts of single thymines not in a T-repeat (top) and in a thymine dinucleotide (bottom); bases are indicated by colours as in panel B. **(D)** Proposed model for adduct formation leading to the mutation patterns of SBSs and indels in A^{n}T. **1**) Adducts are formed on an adenine either inside an adenine-homopolymer (left) or directly adjacent to a thymine-homopolymer (right). **2**) Assuming DNA repair fails to repair the adducted base, during DNA replication, a translesion synthesis polymerase replicates over the adducted base, causing a mismatch or a deletion, represented by **x** in the figure. **3**) In the next cell division the mutation is replicated on the complementary strand. **4**) Following the conventions of the mutational signature field, we display mutations as occurring from the purine of the Watson-Crick base pair.

Fig. 3: **Discovery of SBS_A^{n}T in publicly available mutation data. (A)** Tumours found to be positive for SBS_A^{n}T mutagenesis. **(B)** Example spectra of a bile duct, bladder, HNSCC, prostate and rectal tumour positive for SBS_A^{n}T. The right panel shows the transcriptional strand bias (U = Untranscribed strand, T = transcribed strand). For the whole-genome samples, only SBSs in transcribed regions were included for the strand bias plot. The bile duct, bladder and HNSCC tumours are whole-exome data, the prostate and rectal tumours are whole-genome data.

References

1. Alexandrov LB, Kim L, Haradhvala NL, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, Lopez-Bigas N, Klimczak LJ, McPherson JR, et al. The Repertoire of Mutational Signatures in Human Cancer. *BioRxiv* 2018.

2. Mimaki S, Totsuka Y, Suzuki Y, Nakai C, Goto M, Kojima M, Arakawa H, Takemura S, Tanaka S, Marubashi S, Kinoshita M, Matsuda T, et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. *Carcinogenesis* 2016;**37**: 817-26.

3. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**: D777-D83.

4. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;**136**: E359-86.

5. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**: 2114-20.

6. Boot A, Huang MN, Ng AWT, Ho SC, Lim JQ, Kawakami Y, Chayama K, Teh BT, Nakagawa H, Rozen SG. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res* 2018;**28**: 654-65.

7. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 2017;**171**: 1611-24 e24.

8. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* 2018;**6**: 271-81 e7.

9. Ng AWT, Poon SL, Huang MN, Lim JQ, Boot A, Yu W, Suzuki Y, Thangaraju S, Ng CCY, Tan P, Pang ST, Huang HY, et al. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med* 2017;**9**.

10. Vettore AL, Ramnarayanan K, Poore G, Lim K, Ong CK, Huang KK, Leong HS, Chong FT, Lim TK, Lim WK, Cutcutache I, McPherson JR, et al. Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance. *Genome Med* 2015;**7**: 98.

11. Conaway JW, Conaway RC. Transcription elongation and human disease. *Annu Rev Biochem* 1999;**68**: 301-19.

12. Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng AWT, Boot A, Abedi-Ardekani B, Villar S, Myint SS, Othman R, Poon SL, et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res* 2017;**27**: 1475-86.

13. Sherborne AL, Davidson PR, Yu K, Nakamura AO, Rashid M, Nakamura JL. Mutational Analysis of Ionizing Radiation Induced Neoplasms. *Cell Rep* 2015;**12**: 1915-26.

14. Ichimura M, Ogawa T, Katsumata S, Takahashi K, Takahashi I, Nakano H. Duocarmycins, new antitumor antibiotics produced by Streptomyces; producing organisms and improved production. *J Antibiot (Tokyo)* 1991;**44**: 1045-53.

15. Hurley LH, Rokem JS. Biosynthesis of the antitumor antibiotic CC-1065 by Streptomyces zelensis. *J Antibiot (Tokyo)* 1983;**36**: 383-90.

16. Schwartz GH, Patnaik A, Hammond LA, Rizzo J, Berg K, Von Hoff DD, Rowinsky EK. A phase I study of bizelesin, a highly potent and selective DNA-interactive agent, in patients with advanced solid malignancies. *Ann Oncol* 2003;**14**: 775-82.

17. Cristofanilli M, Bryan WJ, Miller LL, Chang AY, Gradishar WJ, Kufe DW, Hortobagyi GN. Phase II study of adozelesin in untreated metastatic breast cancer. *Anticancer Drugs* 1998;**9**: 779-82.

18. Pavlidis N, Aamdal S, Awada A, Calvert H, Fumoleau P, Sorio R, Punt C, Verweij J, van Oosterom A, Morant R, Wanders J, Hanauske AR. Carzelesin phase II study in advanced breast, ovarian, colorectal, gastric, head and neck cancer, non-Hodgkin's lymphoma and malignant melanoma: a study of the EORTC early clinical studies group (ECSG). *Cancer Chemother Pharmacol* 2000;**46**: 167-71.

19. Kim EG, Kim KM. Strategies and Advancement in Antibody-Drug Conjugate Optimization for Targeted Cancer Therapeutics. *Biomol Ther (Seoul)* 2015;**23**: 493-509.

20. Menderes G, Bonazzoli E, Bellone S, Black J, Altwerger G, Masserdotti A, Pettinella F, Zammataro L, Buza N, Hui P, Wong S, Litkouhi B, et al. SYD985, a novel duocarmycin-based HER2-targeting antibody-drug conjugate, shows promising antitumor activity in epithelial ovarian carcinoma with HER2/Neu expression. *Gynecol Oncol* 2017;**146**: 179-86.

21. Reynolds VL, Molineux IJ, Kaplan DJ, Swenson DH, Hurley LH. Reaction of the antitumor antibiotic CC-1065 with DNA. Location of the site of thermally induced strand breakage and analysis of DNA sequence specificity. *Biochemistry* 1985;**24**: 6228-37.

22. Woynarowski JM. Targeting critical regions in genomic DNA with AT-specific anticancer drugs. *Biochim Biophys Acta* 2002;**1587**: 300-8.

23. Baraldi PG, Cacciari B, Guiotto A, Romagnoli R, Zaid AN, Spalluto G. DNA minor-groove binders: results and design of new antitumor agents. *Farmaco* 1999;**54**: 15-25.

24. Seipke RF, Kaltenpoth M, Hutchings MI. Streptomyces as symbionts: an emerging and widespread theme? *FEMS Microbiol Rev* 2012;**36**: 862-76.

Figure 1

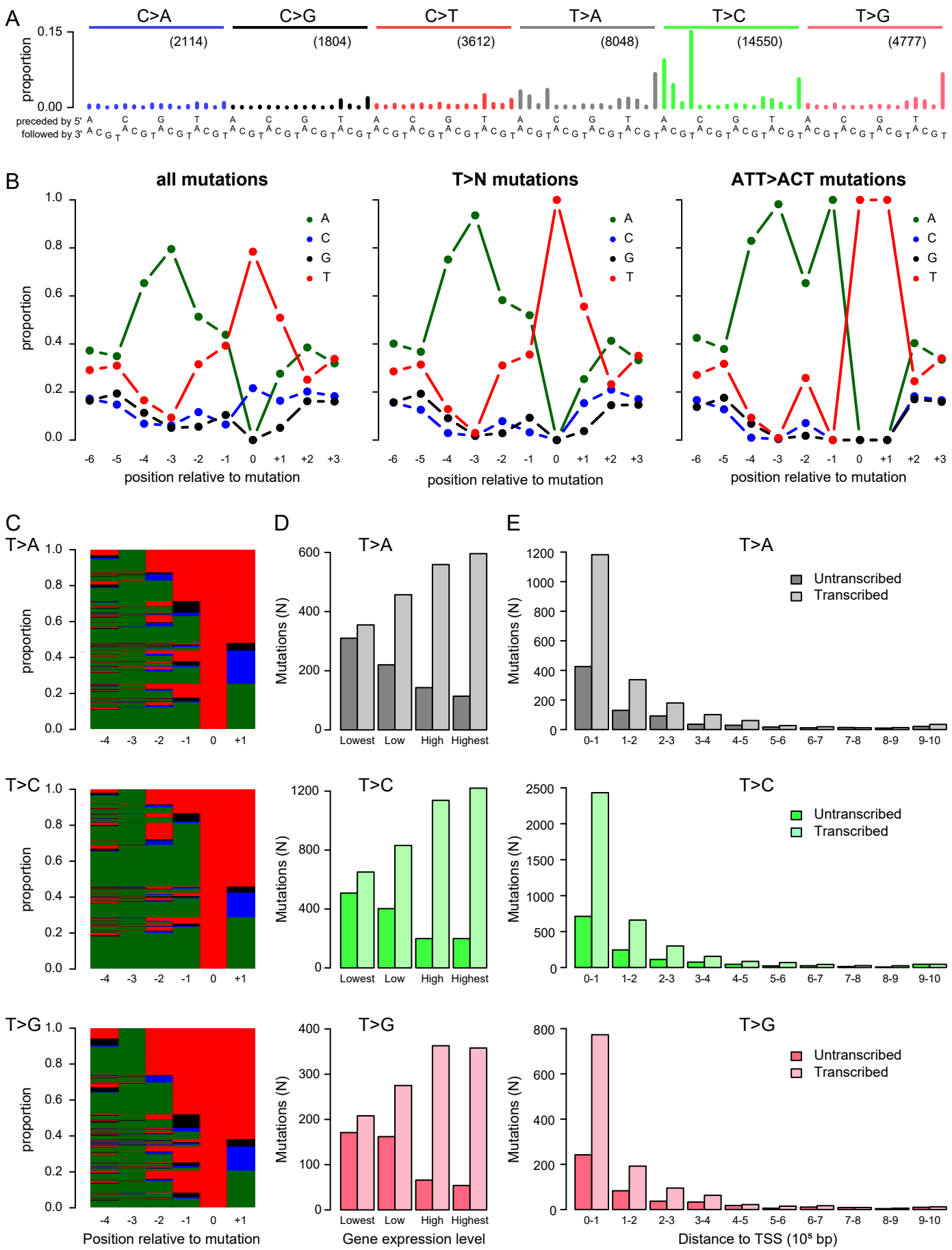
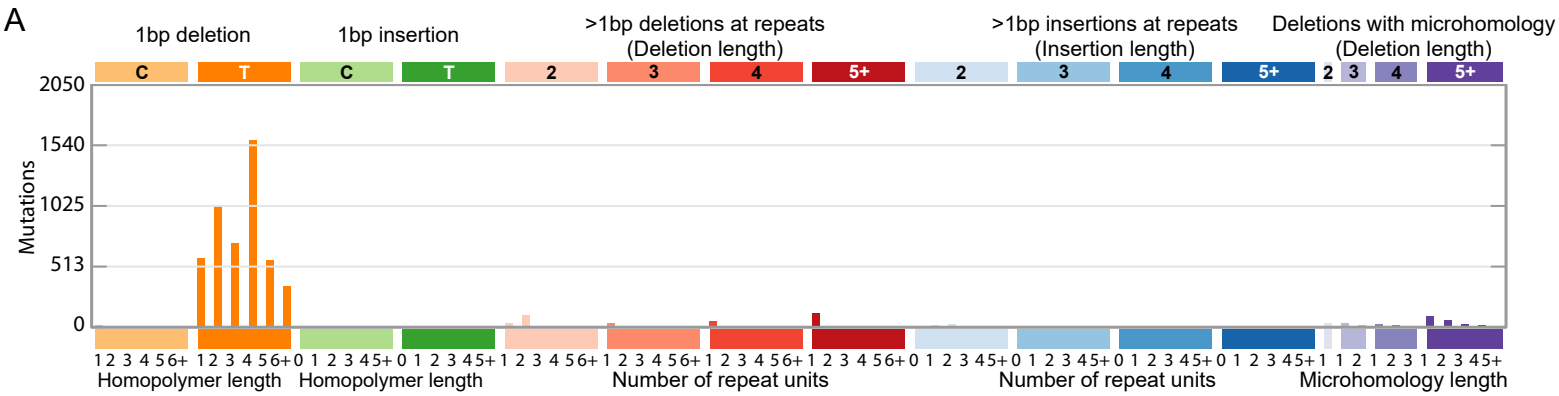
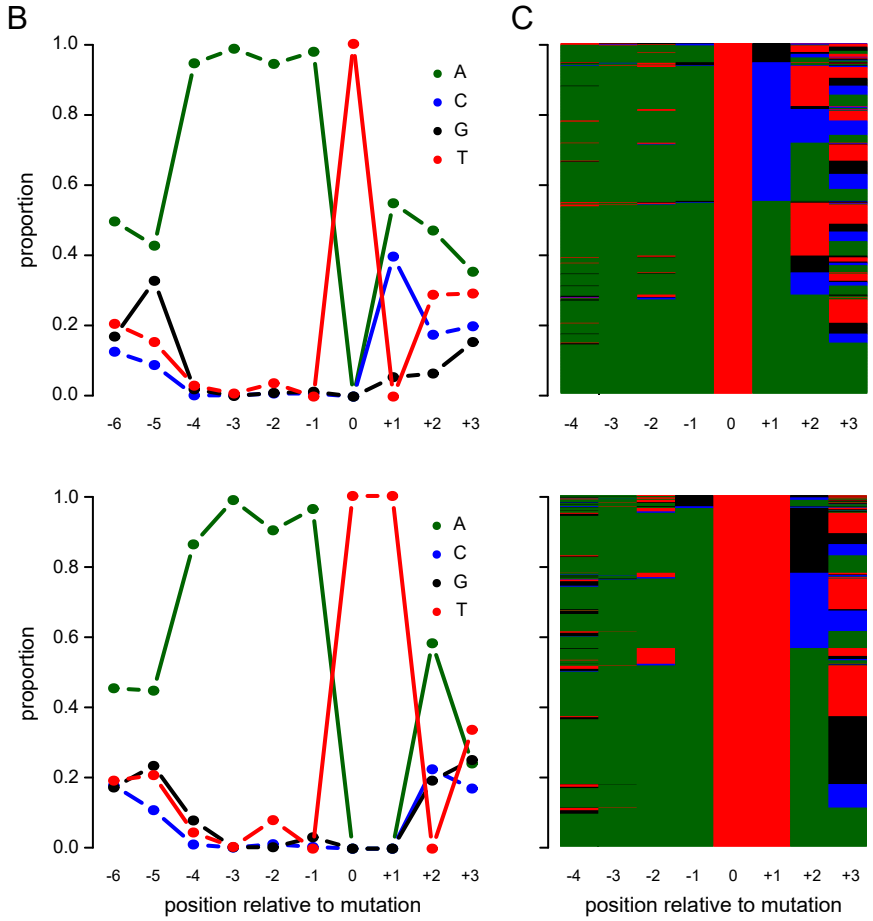

A
C>A
C>G
C>T
T>A
T>C
T>G
(2114)
(1804)
(3612)
(8048)
(14550)
(4777)
proportion
0.15
0.00
preceded by 5'
followed by 3'
B
all mutations
T>N mutations
ATT>ACT mutations
A
C
G
T
proportion
position relative to mutation
C
D
E
T>A
T>C
T>G
Mutations (N)
Lowest
Low
High
Highest
Untranscribed
Transcribed
Position relative to mutation
Gene expression level
Distance to TSS (10⁵ bp)

Figure 2





D

1) Potential adducts

2) Translesion synthesis

3) Further cell divisions

4) Displayed as

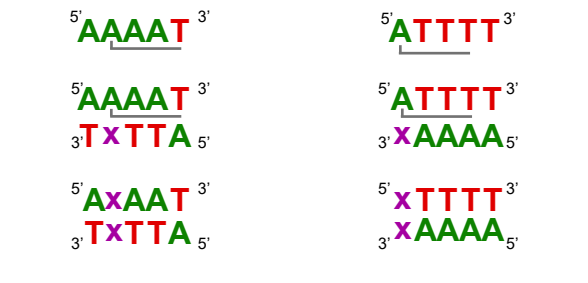


Figure 3

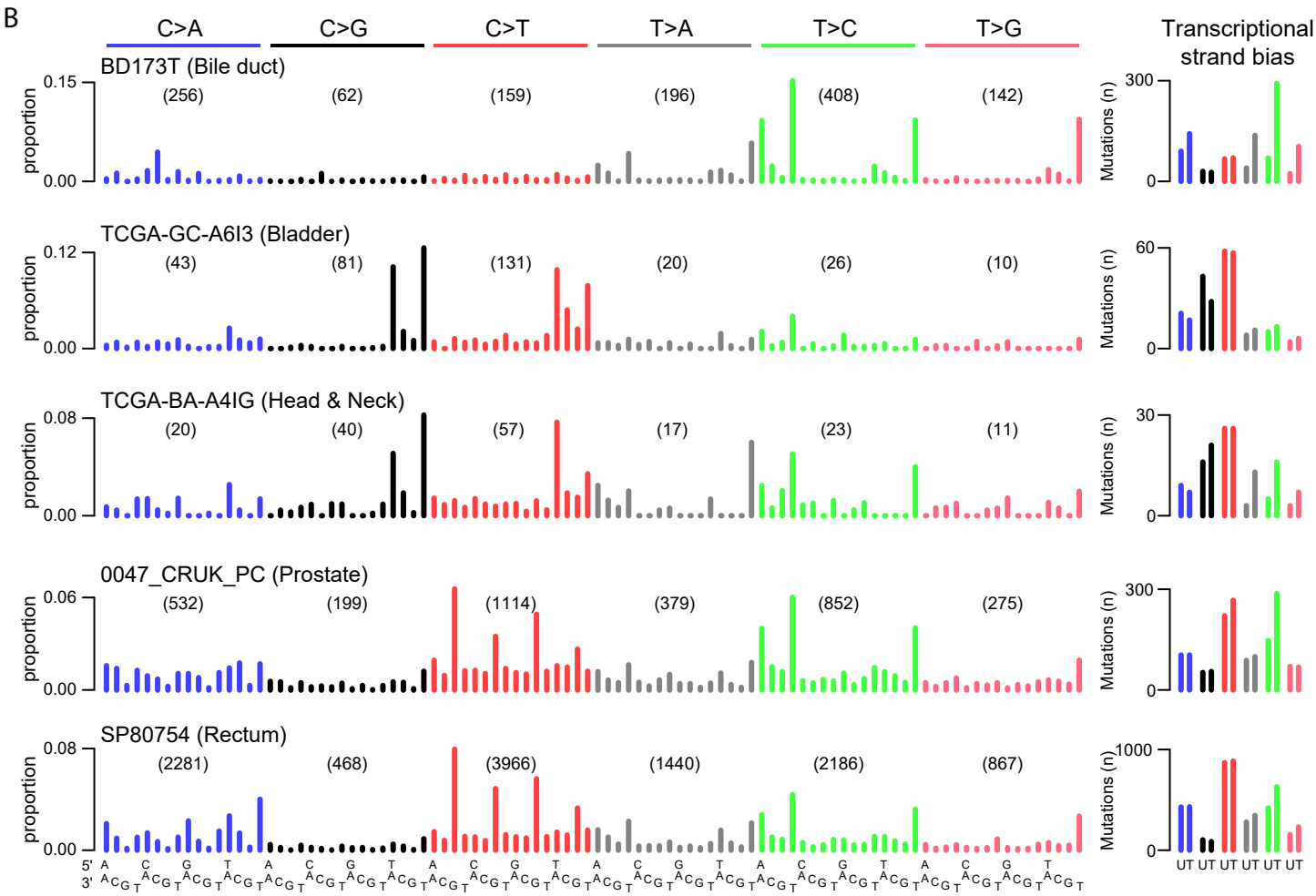
A

Table 1: Whole-exome and whole-genome sequenced tumors found to have significant evidence of SBS_AnT exposure using mSigAct*,**.

Sample	Tissue	Data	1	5	10a	10b	17a	17b	28	Other***	SBS_A^{n}T	mSigAct_qval
BD182T	Bile-duct	WES	349	879	NA	NA	NA	NA	0	NA	877	1.77E-69
BD173T	Bile-duct	WES	0	377	NA	NA	NA	NA	NA	0	846	6.48E-39
BD223T	Bile-duct	WES	0	160	NA	NA	NA	NA	NA	80	182	2.40E-16
BD121T	Bile-duct	WES	117	145	1119	833	NA	NA	NA	NA	70	3.27E-03
TCGA-G2-AA3B	Bladder	WES	0	101	NA	NA	NA	NA	NA	791	45	3.25E-04
TCGA-GC-A6I3	Bladder	WES	12	105	NA	NA	NA	NA	NA	161	32	3.89E-03
TCGA-FU-A3HZ	Cervix	WES	0	107	1418	1003	NA	NA	655	0	71	1.61E-03
TCGA-AY-4071	Colon	WES	45	0	NA	NA	NA	NA	NA	79	40	7.69E-05
sysucc-311T	Colon	WES	0	745	3913	7406	NA	NA	1864	NA	231	8.01E-04
SP80754	Rectum	WGS	2305	2948	NA	NA	NA	NA	206	2097	3653	6.48E-39
SP81711	Rectum	WGS	3789	4065	NA	NA	NA	NA	NA	8512	3908	1.08E-12
TCGA-AG-3902	Rectum	WES	42	41	NA	NA	NA	NA	NA	28	65	1.58E-07
SP81494	Rectum	WGS	2341	12182	NA	NA	NA	NA	NA	10097	2148	7.18E-04
SP80615	Rectum	WGS	0	57839	1070130	795814	NA	NA	479938	NA	22748	2.38E-02
TCGA-DF-A2KV	Endometrium	WES	38	74	1008	938	NA	NA	286	NA	47	2.34E-02
TCGA-QF-A5YS	Endometrium	WES	115	110	1152	1141	NA	NA	220	NA	50	2.38E-02
TCGA-BK-A6W3	Endometrium	WES	123	335	4032	1867	NA	NA	722	NA	90	4.80E-02
LP6005935	Esophagus	WGS	899	11769	NA	NA	5759	16897	NA	NA	9577	3.63E-18
SP111026	Esophagus	WGS	1481	8490	NA	NA	9010	24066	NA	NA	3742	2.77E-10
SP111101	Esophagus	WGS	3450	3123	NA	NA	1011	1607	NA	5476	2245	4.04E-07
TCGA-IG-A4QS	Esophagus	WES	69	112	NA	NA	38	64	NA	NA	36	2.08E-02
TCGA-BA-A4IG	Head & Neck	WES	0	81	NA	NA	NA	NA	NA	43	44	1.51E-06
8069334	Pancreas	WGS	1153	3229	NA	NA	500	1357	NA	NA	862	3.01E-07
PCSI_0060_Pa_X	Pancreas	WES	33	0	NA	NA	NA	NA	NA	32	47	6.68E-07
0047_CRUK_PC	Prostate	WGS	496	1904	NA	NA	NA	NA	NA	NA	951	1.85E-16

*: Displaying only mutational signatures present in at least 5 tumors, the full table is included in Supporting Information Data S2.

**: NA means not analysed, this indicates that this mutational signature was not previously identified in this tumor and therefore was not included in the mSigAct analysis.

***: Total number of SBSs assigned to other mutational signatures.

B