Article

The repertoire of mutational signatures in human cancer

https://doi.org/10.1038/s41586-020-1943-3

Received: 18 May 2018

Accepted: 18 November 2019

Published online: 5 February 2020

Open access

Ludmil B. Alexandrov[1,920], Jaegil Kim[2,920], Nicholas J. Haradhvala[2,3,920], Mi Ni Huang[4,5,920], Alvin Wei Tian Ng[4,5], Yang Wu[4,5], Arnoud Boot[4,5], Kyle R. Covington[6,7], Dmitry A. Gordenin[8], Erik N. Bergstrom[1], S. M. Ashiqul Islam[1], Nuria Lopez-Bigas[9,10,11], Leszek J. Klimczak[12], John R. McPherson[4,5], Sandro Morganella[13], Radhakrishnan Sabarinathan[10,14,15], David A. Wheeler[6,16], Ville Mustonen[17,18,19], PCAWG Mutational Signatures Working Group[20], Gad Getz[2,3,21,22,921], Steven G. Rozen[4,5,23,921]*, Michael R. Stratton[13,921]* & PCAWG Consortium[24]

Somatic mutations in cancer genomes are caused by multiple mutational processes, each of which generates a characteristic mutational signature[1]. Here, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium[2] of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), we characterized mutational signatures using 84,729,690 somatic mutations from 4,645 whole-genome and 19,184 exome sequences that encompass most types of cancer. We identified 49 single-base-substitution, 11 doublet-base-substitution, 4 clustered-base-substitution and 17 small insertion-and-deletion signatures. The substantial size of our dataset, compared with previous analyses[3–15], enabled the discovery of new signatures, the separation of overlapping signatures and the decomposition of signatures into components that may represent associated—but distinct—DNA damage, repair and/or replication mechanisms. By estimating the contribution of each signature to the mutational catalogues of individual cancer genomes, we revealed associations of signatures to exogenous or endogenous exposures, as well as to defective DNA-maintenance processes. However, many signatures are of unknown cause. This analysis provides a systematic perspective on the repertoire of mutational processes that contribute to the development of human cancer.

Somatic mutations in cancer genomes are caused by mutational processes of both exogenous and endogenous origin that operate during the cell lineage between the fertilized egg and the cancer cell[16]. Each mutational process may involve components of DNA damage or modification, DNA repair and DNA replication (which may be normal or abnormal), and generates a characteristic mutational signature that potentially includes base substitutions, small insertions and deletions (indels), genome rearrangements and chromosome copy-number changes[1]. The mutations in an individual cancer genome may have been generated by multiple mutational processes, and thus incorporate multiple superimposed mutational signatures. Therefore, to systematically characterize the mutational processes that contribute to cancer, mathematical methods have previously been used to decipher mutational signatures from somatic mutation catalogues, estimate the number of mutations that are attributable to each signature in individual samples and annotate each mutation class in each tumour with the probability that it arose from each signature[6,9,17–27].

Previous studies of multiple types of cancer have identified more than 30 single-base substitution (SBS) signatures, some of known—but many of unknown—aetiologies, some ubiquitous and others rare, some part of normal cell biology and others associated with abnormal exposures or neoplastic progression[3–5,7–15]. Genome rearrangement signatures have also previously been described[11,25,28–30]. However, the analysis of other classes of mutation has been relatively limited[3,11,31–33].

[1]Department of Cellular and Molecular Medicine, Department of Bioengineering, Moores Cancer Center, University of California, San Diego, CA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. [4]Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore. [5]Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore. [6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [7]Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA. [8]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [9]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [10]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. [11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [12]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [13]Wellcome Sanger Institute, Hinxton, UK. [14]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India. [15]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [16]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [17]Department of Computer Science, University of Helsinki, Helsinki, Finland. [18]Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland. [19]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [20]A list of members and their affiliations appears at the end of the paper. [21]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [22]Harvard Medical School, Boston, MA, USA. [23]SingHealth, Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore, Singapore. [24]A list of members and their affiliations appears online. [920]These authors contributed equally: Ludmil B. Alexandrov, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang. [921]These authors jointly supervised this work: Gad Getz, Steven G. Rozen, Michael R. Stratton. *e-mail: steverozen@gmail.com; mrs@sanger.ac.uk

Mutational signature analysis has predominantly used cancer exome sequences. However, the many-fold-greater numbers of somatic mutations in whole genomes provide substantially increased power for signature decomposition, enabling the better separation of partially correlated signatures and the extraction of signatures that contribute relatively small numbers of mutations. Furthermore, technical artefacts and differences in sequencing technologies and mutation-calling algorithms can themselves generate mutational signatures. Therefore, the uniformly processed and highly curated sets of all classes of somatic mutations from the 2,780 cancer genomes of the PCAWG project[2], combined with most other suitable cancer genomes (accession code syn11801889, available at https://www.synapse.org/#!Synapse:syn11801889), present a notable opportunity to establish the repertoire of mutational signatures and determine their activities across different types of cancer. The timing of these signatures during the evolution of individual cancers and the repertoire of signatures of structural variation have been explored in other PCAWG analyses[30,34].

Mutational signature analysis

The 23,829 samples—which include most types of cancer, and comprise the 2,780 PCAWG whole genomes[2], 1,865 additional whole genomes and 19,184 exomes—yielded 79,793,266 somatic SBSs, 814,191 doublet-base substitutions (DBSs) and 4,122,233 small indels that were analysed for mutational signatures, about 10-fold-more mutations than any previous study of which we are aware (syn11801889)[6].

We developed classifications for each type of mutation. For SBSs, the primary classification comprised 96 classes (available at https://cancer.sanger.ac.uk/cosmic/signatures/SBS) constituted by the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G (in which the mutated base is represented by the pyrimidine of the base pair), plus the flanking 5′ and 3′ bases. In some analyses, two flanking bases 5′ and 3′ to the mutated base were considered (producing 1,536 classes) or mutations within transcribed genome regions were selected and classified according to whether the mutated pyrimidine fell on the transcribed or untranscribed strand (producing 192 classes). We also derived a classification for DBSs (78 classes; available at https://cancer.sanger.ac.uk/cosmic/signatures/DBS). Indels were classified as deletions or insertions and—when of a single base—as C or T, and according to the length of the mononucleotide repeat tract in which they occurred. Longer indels were classified as occurring at repeats or with overlapping microhomology at deletion boundaries, and according to the size of indel, repeat and microhomology (83 classes; available at https://cancer.sanger.ac.uk/cosmic/signatures/ID).

The PCAWG whole-genome sequences, the additional whole-genome sequences and the exome sequences were each analysed separately (syn11801889)[2]. Signatures were extracted from each type of cancer individually, from all cancer types together, as separate SBS, DBS and indel signatures, and as composite signatures of all three types of mutation (Supplementary Note 2).

We used two methods based on nonnegative matrix factorization (NMF): SigProfiler, an elaborated version of the framework used for the previous 'Catalogue Of Somatic Mutations In Cancer' (COSMIC) compendium of mutational signatures (COSMIC v.2, available at https://cancer.sanger.ac.uk/cosmic/signatures_v2)[11,17], and SignatureAnalyzer, which is based on a Bayesian variant of NMF[9,27,35]. NMF determines the signature profiles and contributions of each signature to each cancer genome as part of its factorization of the input matrix of mutation spectra. However, with many signatures and/or heterogeneous mutation burdens across samples, the mutations observed in a particular sample can be reconstructed in multiple ways—often with small and/or biologically implausible contributions from many signatures. Therefore, each method has developed a separate procedure for estimating the contributions of signatures to each sample (Methods).

We tested SignatureAnalyzer and SigProfiler on 11 sets of synthetic data (including 64,400 synthetic samples), generated from known signature profiles (Methods, Supplementary Note 2). Both methods performed well in re-extracting known signatures from realistically complex data. Extracted signatures that were discordant from the known input usually arose from difficulties in selecting the correct number of signatures. The results confirm that use of NMF-based approaches for extracting mutational signatures is not a purely algorithmic process, but also requires consideration of evidence from experimentally determined mutational signatures and the DNA damage and repair literature, prior evidence of biological plausibility and human-guided sensitivity analysis confirming that extractions from different groupings of tumours yield consistent results. We used these types of evidence and approaches in determining the signature profiles reported here. The findings are consistent with results regarding NMF, and the related areas of probabilistic topic modelling and latent Dirichlet allocation, in multiple problem domains[36,37]. It is widely understood that the choice of the number of latent variables (for our purposes, the number of mutational signatures) is rarely amenable to complete automation.

The results from our SigProfiler and SignatureAnalyzer analyses of cancer data exhibited many similarities, and we assigned the same identifiers to similar signatures extracted using the two methods (syn12016215). However, there were also noteworthy differences. The numbers of SBS signatures found in PCAWG tumours with a low mutation burden (94.4% of cases that contain 47% of mutations) were similar: 31 using SigProfiler and 35 using SignatureAnalyzer. However, the numbers of additional SBS signatures extracted from hypermutated PCAWG samples (5.6% of cases, containing 53% of mutations) were different: 13 using SigProfiler and 25 using SignatureAnalyzer. There were also differences in SBS signature profiles, including among signatures found in cases with a low mutation burden. The latter primarily involved relatively featureless ('flat') signatures, which are mathematically challenging to deconvolute. Finally, there were differences in signature attributions to individual samples. SignatureAnalyzer used more signatures to reconstruct the mutational profiles (Extended Data Fig. 1) (syn12169204 and syn12177011) and attributions to flat signatures were different (Extended Data Fig. 2a, b) (syn12169204). The DBS and indel signatures were generally similar between the two methods (Extended Data Fig. 2c, d).

The final reference mutational signatures were determined from the PCAWG set, supplemented by additional signatures from the other datasets (COSMIC, available at https://cancer.sanger.ac.uk/cosmic/signatures). Each signature was allocated an identifier consistent with, and extending, the COSMIC v.2 annotation. Some previous signatures split into multiple constituent signatures: these were numbered as in the previous annotation, but with additional letter suffixes (for example, SBS17 was split into SBS17a and SBS17b). DNA sequencing and analysis artefacts also generate mutational signatures. We indicate which signatures are possible artefacts but do not present them below (full information is available at https://cancer.sanger.ac.uk/cosmic/signatures). The results of both SignatureAnalyzer and SigProfiler were used throughout the study. However, for brevity and for continuity with the signature set previously displayed in COSMIC v.2—which has been widely used as a reference—SigProfiler results are outlined here, and SignatureAnalyzer results are provided in Extended Data Figs. 3, 4 and at syn11738307.

Single-base substitution signatures

There were substantial differences in the numbers of SBSs between samples (ranging from hundreds to millions) and between cancer types[38] (Fig. 1). In total, 67 SBS mutational signatures were extracted, of which 49 were considered likely to be of biological origin (Fig. 2, Methods; available at https://cancer.sanger.ac.uk/cosmic/signatures/SBS/). Except for signature SBS25, all signatures reported in COSMIC v.2 (ref. [6]) were confirmed; the median cosine similarity between the newly

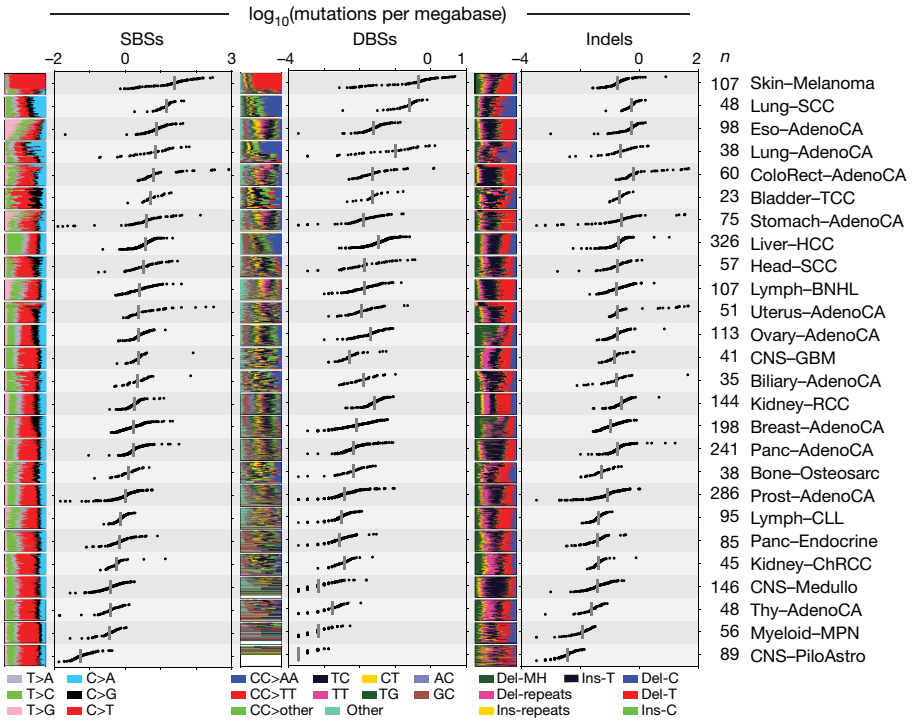


Fig. 1 | Mutation burdens of SBSs, DBSs and small indels across PCAWG tumour types. The numbers of cases of each tumour type are shown next to the labels. Each dot represents one tumour. Tumour types are ordered by the median numbers of single-base substitutions. Only tumour types with >20 samples are shown. AdenoCA, adenocarcinoma; BNHL, B-cell non-Hodgkin lymphoma; ChRCC, chromophobe renal cell carcinoma; CLL, chronic lymphocytic leukaemia; CNS, central nervous system; ColoRect, colorectal; Eso, oesophageal; GBM, glioblastoma; HCC, hepatocellular carcinoma; Medullo, medulloblastoma; MH, microhomology; MPN, myeloproliferative neoplasm; Osteosarc, osteosarcoma; Panc, pancreatic; PiloAstro, pilocytic astrocytoma; Prost; prostate; RCC, renal cell carcinoma; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; Thy, thyroid.

derived signatures and those on COSMIC v.2 was 0.95, excluding the 'split' signatures (discussed below). SBS25 was previously found in cell lines derived from Hodgkin lymphomas treated with chemotherapy, and no primary cancers of this type were available. The newly derived signatures showed much improved separation from each other and more-distinct signature profiles, as compared with COSMIC v.2 signatures (see 'Better separation compared to COSMIC v.2 signatures' in Supplementary Note 2 for more information).

Thirteen of the SBS signatures we extracted (excluding those due to signature splitting) represent newly identified and probably real signatures, not present in COSMIC v.2. Some were rare (SBS31, SBS32, SBS35, SBS36, SBS42 and SBS44). Others were more common, but contributed relatively few mutations and/or were similar to previously discovered signatures (SBS38, SBS39 and SBS40). Notably, SBS40 is a flat signature similar to SBS5. It contributes to multiple types of cancer, but its similarity to SBS5 renders the extent of this contribution uncertain. For some of the newly identified signatures, there were plausible underlying aetiologies (Fig. 3, Extended Data Figs. 4, 5): for SBS31 and SBS35, platinum compound chemotherapy[39]; for SBS32, azathioprine therapy; for SBS36, inactivating germline or somatic mutations in *MUTYH* (which encodes a component of the base excision repair machinery)[40,41]; for SBS38, additional effects of exposure to ultraviolet (UV) light; for SBS42, occupational exposure to haloalkanes[13]; and for SBS44, defective DNA mismatch repair[42].

Three previously characterized base substitution signatures (SBS7, SBS10 and SBS17) split into multiple constituent signatures (Fig. 2). Signature splitting probably reflects the existence of multiple distinct mutational processes initiated by the same exposure that have closely—but not perfectly—correlated activities. We previously regarded SBS7 as a single signature composed predominantly of C>T at C<u>C</u>N and T<u>C</u>N trinucleotides (the mutated base is underlined) together with many fewer T>N mutations. It was found in malignant melanomas and squamous skin carcinomas, and is probably due to the UV-light-induced formation of pyrimidine dimers, followed by translesion DNA synthesis by error-prone polymerases predominantly inserting A opposite to damaged cytosines. SBS7 has now been decomposed into four constituent signatures. SBS7a and SBS7b (consisting mainly of C>T at T<u>C</u>N and C>T at C<u>C</u>N, respectively) may reflect different pyrimidine-dimer photoproducts. SBS7c and SBS7d (consisting predominantly of T>A at N<u>T</u>T and T>C at N<u>T</u>T, respectively[43]) may be due to low frequencies of the misincorporation of T and G opposite to thymines in pyrimidine dimers. The splitting of SBS10 and SBS17 is described at https://cancer.sanger.ac.uk/cosmic/signatures/SBS/.

Several base substitution signatures showed transcriptional strand bias, which may be attributable to transcription-coupled nucleotide excision repair acting on DNA damage and/or to an excess of DNA damage on untranscribed strands of genes[44]. Both mechanisms result in more mutations of damaged bases on untranscribed than on transcribed strands of genes. Assuming that either mechanism is responsible for the observed transcriptional strand biases, DNA damage to cytosine (SBS7a and SBS7b), guanine (SBS4, SBS8, SBS19, SBS23, SBS24, SBS31, SBS32, SBS35 and SBS42), thymine (SBS7c, SBS7d, SBS21, SBS26 and SBS33) and adenine (SBS5, SBS12, SBS16, SBS22 and SBS25) may underlie these mutational signatures (plots of strand bias are available at https://cancer.sanger.ac.uk/cosmic/signatures/SBS/). The likely DNA-damaging agents are known for SBS4 (tobacco mutagens), SBS7a, SBS7b, SBS7c and SBS7d (UV light), SBS22 (aristolochic acid), SBS24 (aflatoxin), SBS25 (chemotherapy), SBS31 and SBS35 (platinum compounds), SBS32 (azathioprine) and SBS42 (haloalkanes).

Using the SBS classification of 1,536 mutation types, which uses the sequence context two bases 5′ and two bases 3′ to each mutated base, yielded signatures that are largely consistent with those based on substitutions in trinucleotide contexts. Notably, however, two forms of both SBS2 and SBS13 were extracted, one with mainly a pyrimidine and the other with mainly a purine at the −2 base (the second base 5′ to the mutated cytosine). These may represent the activities of the cytidine deaminases APOBEC3A and APOBEC3B, respectively[45]. If so, APOBEC3A accounts for many more mutations than APOBEC3B in cancers with high APOBEC activity. Other signatures showed nonrandom sequence contexts at +2 and −2 positions (for example, SBS17a, SBS17b and SBS9), but sequence context effects were generally much stronger for bases immediately 5′ and 3′ to mutated bases.

SBS signatures showed substantial variation in the numbers of cancer types and cancer samples in which they were found, and in the mutations attributed per cancer sample (Fig. 3). Almost all individual cancer samples exhibited multiple signatures, with a mode of three in

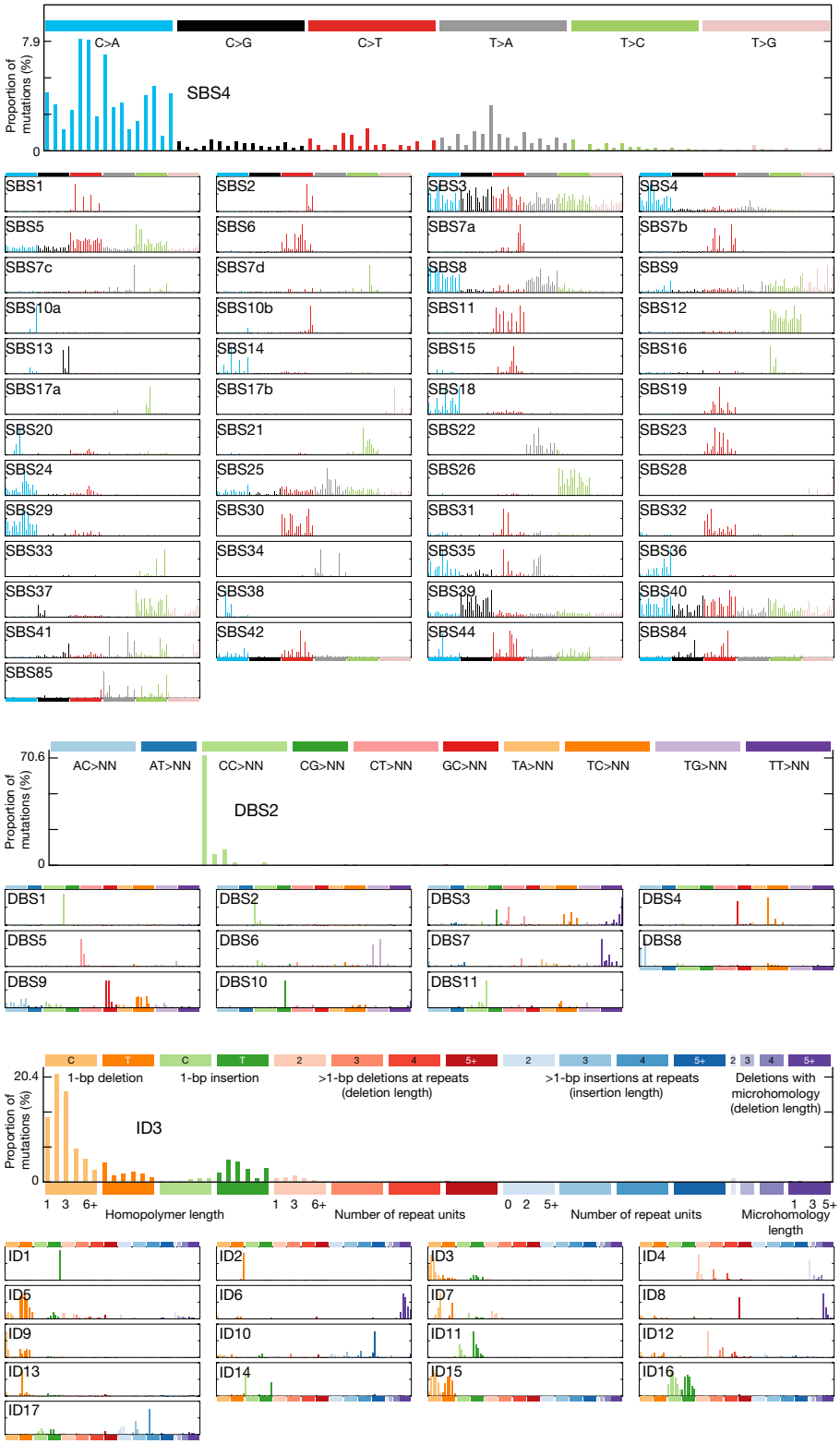


Fig. 2 | Profiles of SBS, DBS and small indel mutational signatures. The classifications of each mutation type (SBS, 96 classes; DBS, 78 classes; and indels, 83 classes) are described in the main text. Magnified versions of signatures SBS4, DBS2 and ID3 (all of which are associated with tobacco smoking) are shown to illustrate the positions of each mutation subtype on each plot. The plotted data are available in digital form (along with the *x* axis labels) at syn12025148.

the PCAWG set (syn12169204). The assigned signatures reconstruct well the mutational spectra of the cancer samples (in PCAWG samples, the median cosine similarity was 0.97; 96.3% of samples with cosine similarity >0.90): Fig. 4 shows illustrative examples.

Some mutational processes generate base substitutions that cluster in small genomic regions. The limited numbers of such mutations may result in a failure to detect their signatures using standard methods. We therefore identified clustered mutations in each genome and analysed

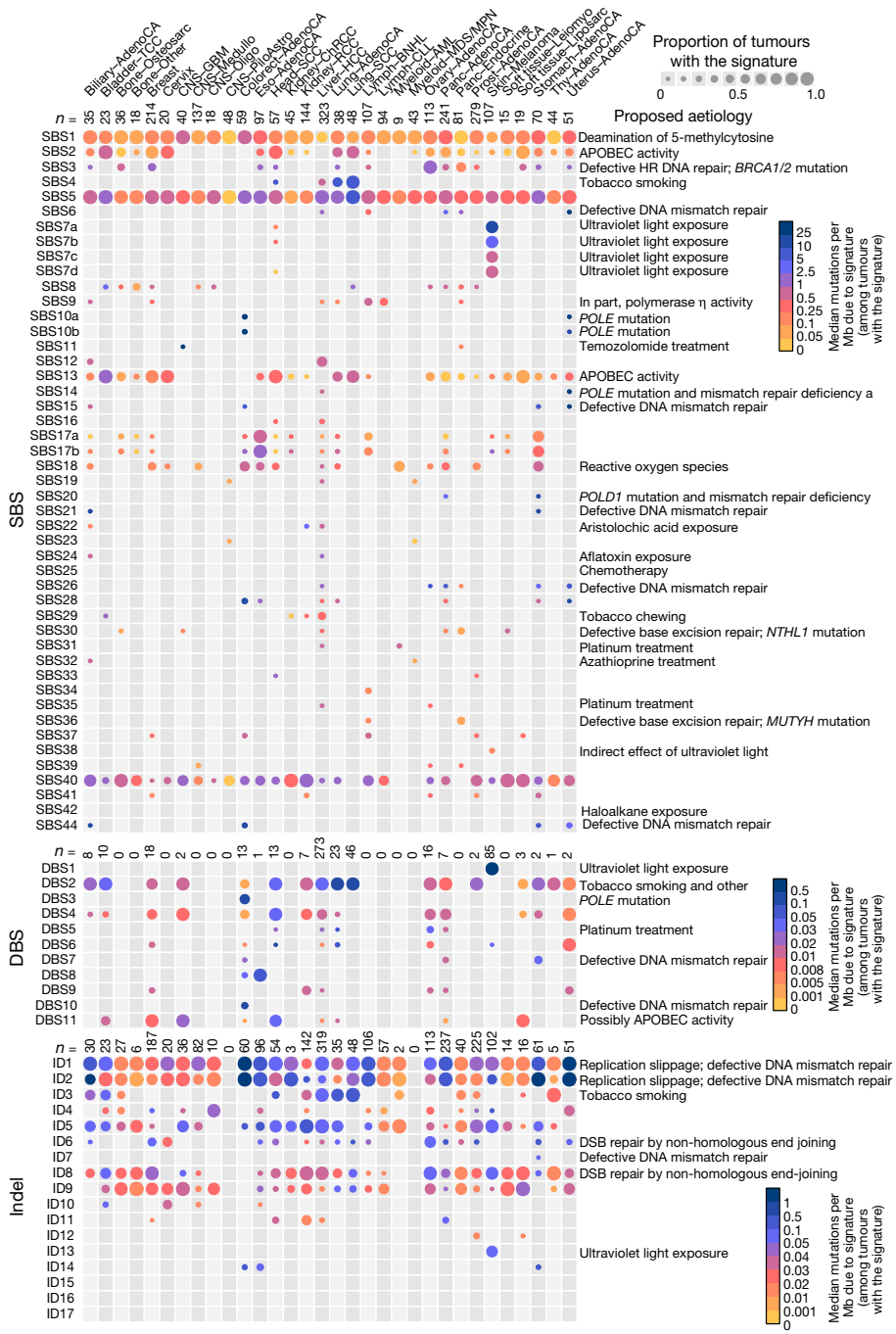


Fig. 3 | The number of mutations contributed by each mutational signature to the PCAWG tumours. The size of each dot represents the proportion of samples of each tumour type that shows the mutational signature. The colour of each dot represents the median mutation burden of the signature in samples that show the signature. Tumours that had few mutations or that were poorly reconstructed by the signature assignment were excluded. The contributions of composite signatures to the PCAWG cancers, and SBS signatures to the complete set of cancer samples analysed, are shown in Extended Data Figs. 4 and 5, respectively. AML, acute myeloid leukaemia; liposarc, liposarcoma; MDS, myelodysplastic syndrome.

them separately (Methods). Four main clustered mutational signatures were identified (Fig. 2), as previously reported[4,27,32]. Two, which are found in multiple types of cancer, were similar to SBS2 and SBS13 (which have been attributed to APOBEC enzyme activity) and represent foci of kataegis[3,32,46]. Two further clustered signatures, one characterized by C>T and C>G mutations at (A or G)C(C or T) trinucleotides[47] and the other T>A and T>C mutations at (A or T)T(A or T), were found in lymphoid neoplasms; they probably represent the direct and indirect consequences of activation-induced cytidine deaminase mutagenesis and translesion DNA synthesis by error-prone polymerases (SBS84 and SBS85, respectively)[27].

Doublet-base substitution signatures

Tandem doublet, triplet, quadruplet, quintuplet and sextuplet base substitutions (syn11801938 and syn11726620) were observed at about 1% the prevalence of SBSs. In most cancer genomes, the number of DBSs was considerably higher than would be expected from the random adjacency of SBSs (syn12177057), indicating the existence of commonly occurring, single mutagenic events that cause substitutions at neighbouring bases. There was substantial variation in the number of DBSs, ranging from 0 to 20,818 in a sample. The numbers of DBSs were generally proportional to the numbers of SBSs (Fig. 1), although colorectal adenocarcinomas had fewer than expected, and lung cancers and melanomas had more (Extended Data Table 1). We extracted eleven DBS signatures (Fig. 2, of which three have previously been reported[33,48].

Signature DBS1 was characterized by CC>TT mutations (Fig. 2), contributed hundreds to tens of thousands of mutations in malignant melanomas with SBS7a and SBS7b (Fig. 3), exhibited transcriptional strand bias consistent with damage to cytosines (syn12177063) and is a known consequence of DNA damage induced by UV light[33,49]. Excluding cancers associated with exposure to UV light also yielded a signature (DBS11) that was characterized predominantly by CC>TT mutations, but only

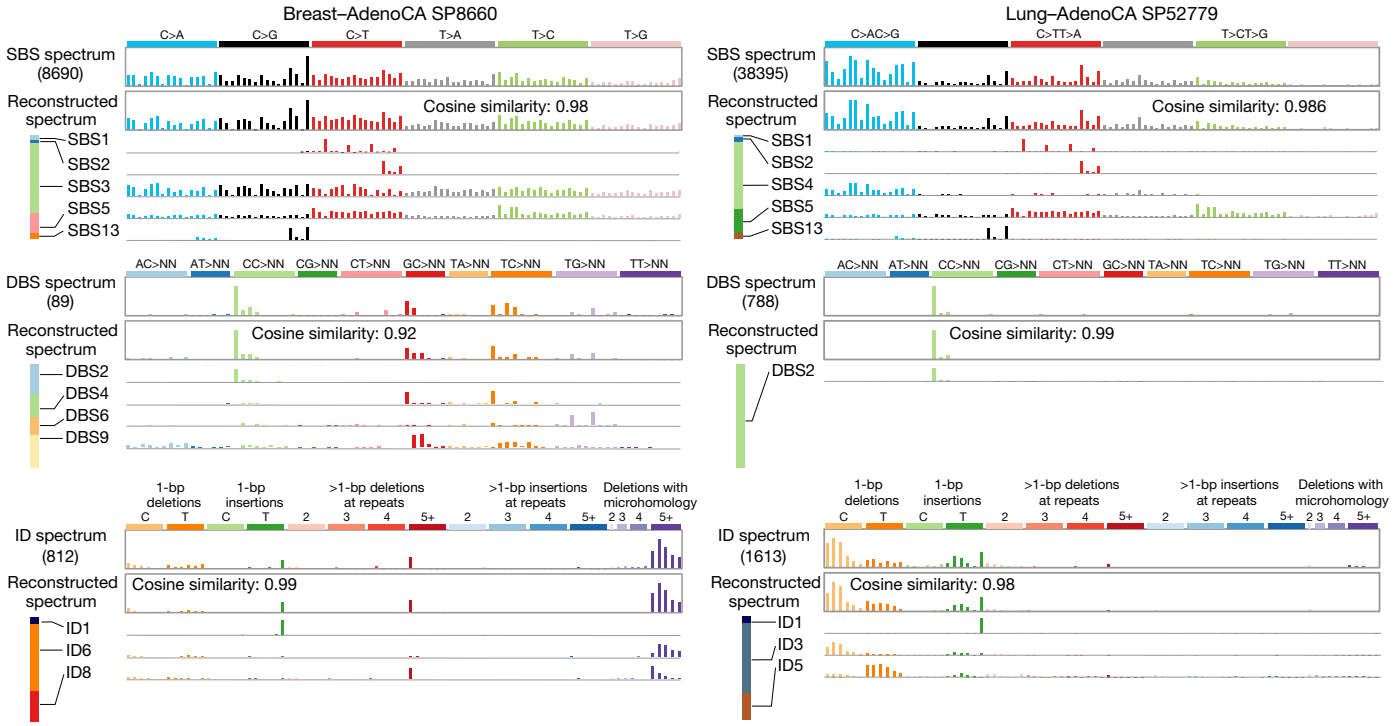


Fig. 4 | Illustrative examples of mutational spectra of individual cancer samples. The contributory SBS, DBS and small indel mutational signatures in two tumours are shown.

contributing tens of mutations in many samples from multiple types of cancer (Figs. 2, 3). DBS11 was associated with SBS2, which is due to APOBEC activity: APOBEC activity may, therefore, also generate DBS11.

DBS2 was composed predominantly of CC>AA mutations, with smaller numbers of CC>AG and CC>AT mutations, and contributed hundreds to thousands of mutations in lung adenocarcinoma, lung squamous and head and neck squamous carcinomas, which are often caused by tobacco smoking[33] (Figs. 2, 3). DBS2 showed transcriptional strand bias indicative of guanine damage (syn12177064) and was associated with SBS4, which is caused by exposure to tobacco smoke. It is likely, therefore, that DBS2 can be a consequence of DNA damage by tobacco-smoke mutagens.

A signature similar to DBS2 contributed hundreds of mutations to liver cancers and tens of mutations to other types of cancer without evidence of exposure to tobacco smoke. A pattern resembling DBS2 also dominates DBSs in healthy mouse cells[50]. The nature of the mutational processes that underlie these signatures in human cancers that are unrelated to smoking, and in healthy mice, is unknown. However, in experimental systems, acetaldehyde exposure has been shown to generate a mutational signature characterized primarily by CC>AA mutations, and lower burdens of CC>AG and CC>AT mutations, together with C>A SBSs[48]. Acetaldehyde is an oxidation product of alcohol and a constituent of cigarette smoke. The role of acetaldehyde, and perhaps other aldehydes, in generating DBS2 merits further investigation[51].

DBS3, DBS7, DBS8 and DBS10 showed hundreds to thousands of mutations in rare colorectal, stomach and oesophageal cancers, some of which showed evidence of defective DNA mismatch repair (DBS7 and DBS10) or polymerase epsilon exonuclease domain mutations (DBS3) that generate hypermutator phenotypes (Figs. 2, 3). DBS5 was found in cancers exposed to platinum chemotherapy, and is associated with SBS31 and SBS35.

Small insertion-and-deletion signatures

Indels were usually present at about 10% of the frequency of base substitutions (Fig. 1). There was substantial variation between cancer genomes in the number of indels, even when cancers with evidence of defective DNA mismatch repair were excluded. Overall, the numbers of deletions and insertions were similar, but there was variation between cancer types: some cancers showed more deletions and others more insertions of various subtypes (Fig. 1). We extracted 17 indel mutational signatures (Fig. 2).

Indel signature 1 (ID1) was composed predominantly of insertions of thymine and ID2 was composed predominantly of deletions of thymine, both at long (≥5) thymine mononucleotide repeats (Fig. 2). Tens to hundreds of mutations of both signatures were found in most samples of most types of cancer, but were particularly common in colorectal, stomach, endometrial and oesophageal cancers and in diffuse large B cell lymphoma (Fig. 3). Together, ID1 and ID2 accounted for 97% and 45% of indels in hypermutated and non-hypermutated cancer genomes, respectively (Extended Data Table 2). They are probably due to slippage of either the nascent (ID1) or template strand (ID2) during DNA replication of long mononucleotide tracts.

ID3 was characterized predominantly by deletions of cytosine at short (≤5-bp long) mononucleotide cytosine repeats and exhibited hundreds of mutations in cancers of the lung, head and neck that are associated with tobacco smoking (Figs. 2, 3). There was transcriptional strand bias of mutations, with more guanine deletions than cytosine deletions on the untranscribed strands of genes, which is compatible with transcription-coupled nucleotide excision repair of damaged guanine (syn12177065 and syn12177066). The numbers of ID3 mutations positively correlated with the numbers of SBS4 and DBS2 mutations, which we have shown are associated with tobacco smoking (Extended Data Figs. 6, 7). Thus, DNA damage by components of tobacco smoke probably underlie ID3.

ID13 was characterized predominantly by deletions of thymine at thymine–thymine dinucleotides and exhibited large numbers of mutations in malignant melanomas of the skin (Figs. 2, 3). The numbers of ID13 mutations correlated with the numbers of SBS7a, SBS7b and DBS1 mutations, which we have attributed to DNA damage induced by UV light (Extended Data Figs. 6, 7). However, deletions of cytosine

at cytosine–cytosine dinucleotides did not feature strongly in ID13, which may reflect the predominance of thymine compared to cytosine dimers induced by UV light[52].

ID6 and ID8 were both characterized predominantly by ≥5-bp deletions (Fig. 2). ID6 exhibited overlapping microhomology at deletion boundaries with a mode of 2 bp (and often longer stretches) and correlated with SBS3, which we have attributed to defective homologous-recombination-based repair (Extended Data Figs. 6, 7). By contrast, ID8 deletions showed shorter or no microhomology at deletion boundaries and did not strongly correlate with SBS3. Both deletion patterns may be characteristic of DNA double-strand-break repair by non-homologous-recombination-based end-joining mechanisms and—if so—this suggests that at least two distinct forms are operative in human cancer[53].

A small fraction of cancers exhibited very large numbers of ID1 and ID2 mutations (>10,000) (Fig. 3) (shown at https://cancer.sanger.ac.uk/cosmic/signatures/ID). These were usually accompanied by SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and/or SBS44, which are associated with deficiency in DNA mismatch repair—sometimes combined with POLE or POLD1 proofreading deficiency (SBS14 and SBS20)[35]. Occasional cases with these signatures additionally showed large numbers of indels attributed to ID7 (syn11738668), and rare samples showed large numbers of ID4, ID11, ID14, ID15, ID16 or ID17 mutations but did not show large numbers of ID1 and ID2 mutations or the SBS signatures associated with deficiency in DNA mismatch repair.

Correlations with age

A positive correlation between age of cancer diagnosis and the number of mutations attributable to a signature suggests that the underlying mutational process has been operative (at a more or less constant rate) throughout the cell lineage from fertilized egg to cancer cell, and thus in the normal cells from which that type of cancer develops[6,54]. Confirming previous reports[6,54], the numbers of SBS1 and SBS5 mutations correlate with age, and exhibit different rates in different types of tissue (*q* values provided in syn12030687, syn20317940 and syn12217988). SBS40 also correlated with age in multiple types of cancer, although—given its similarity to SBS5—misattribution cannot be excluded. DBS2 and DBS4 correlated with age; consistent with activity in normal cells and, when combined their profiles closely resemble the spectrum of DBS mutations found in normal mouse cells[50]. ID1, ID2, ID5 and ID8 showed correlations with age in multiple tissues. ID1 and ID2 indels are probably due to slippage at poly T repeats during DNA replication and correlated with the numbers of SBS1 substitutions, which have previously been proposed to reflect the number of mitoses a cell has experienced[6]. Thus, SBS1, ID1 and ID2 may all be generated during DNA replication at mitosis. The number of ID5 mutations correlated with the number of SBS40 mutations, and the mutational processes that underlie these two age-correlated signatures may therefore contain common components. ID8, which is predominantly composed of ≥5-bp deletions with no or 1 bp of microhomology at their boundaries, is probably due to DNA double-strand breaks repaired by a non-homologous-end-joining mechanism. The results indicate that multiple mutational processes operate in normal cells.

Discussion

There are important constraints, limitations and assumptions in the analytic frameworks used here to characterize mutational signatures. Signatures extracted from sample sets in which multiple processes are operative remain mathematical approximations, with profiles that are potentially influenced by the mathematical approach used and other factors. For conceptual and practical simplicity, we assume that a single signature is associated with each mutational process and provide an average reference signature to represent it. However, we do not discount the possibility that further nuances and variations of signature profiles exist. We have estimated the contributions from each signature to the mutation burden in each sample. However, with increasing numbers of signatures and differences of multiple orders of magnitude in mutation burdens between some signatures, prior knowledge has helped to avoid biologically implausible results. Thus, the further development of methods for deciphering and attributing mutational signatures is warranted, ideally supported by signatures derived from experimental systems in which the causes are known. Nevertheless, signatures with many similarities and some differences can be found by different mathematical approaches, and these can be confirmed in several ways, including experimentally elucidated signatures[5,31,39,42,43,54–62] and tumours dominated by a single signature (syn12016215).

This analysis includes most publicly available exome and whole-genome cancer sequences. Some rare or geographically restricted signatures may not have been captured, signatures conferring limited mutation burdens may have been missed and signatures of therapeutic mutagenic exposures have not been exhaustively explored. Nevertheless, it is likely that a substantial proportion of the naturally occurring mutational signatures found in human cancer have now been described. This comprehensive repertoire provides a foundation for research into the aetiologies of geographical and temporal differences in cancer incidence, the mutational processes that operate in healthy tissues and non-neoplastic disease states, clinical and public health applications of signatures and mechanistic understanding of the mutational processes that underlie carcinogenesis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-1943-3.

1. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
2. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
3. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
4. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
5. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
6. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
7. Poon, S. L. et al. Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.* **7**, 38 (2015).
8. Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
9. Kim, J. et al. Somatic *ERCC2* mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
10. Merlevede, J. et al. Mutation allele burden remains unchanged in chronic myelomonocytic leukaemia responding to hypomethylating agents. *Nat. Commun.* **7**, 10767 (2016).
11. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
12. Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**, 531–540 (2016).
13. Mimaki, S. et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. *Carcinogenesis* **37**, 817–826 (2016).
14. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
15. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
16. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
17. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
18. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).

19. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol*. **14**, R39 (2013).
20. Roberts, N. *hdp (hierarchical Dirichelet process) R package* https://github.com/nicolaroberts/hdp (2015).
21. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
22. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet*. **11**, e1005657 (2015).
23. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).
24. Ardin, M. et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17**, 170 (2016).
25. Funnell, T. et al. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol*. **15**, e1006799 (2019).
26. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. **10**, 33 (2018).
27. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun*. **6**, 8866 (2015).
28. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res*. **23**, 228–235 (2013).
29. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
30. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* https://doi.org/10.1038/s41586-019-1913-9 (2020).
31. Meier, B. et al. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*. **24**, 1624–1636 (2014).
32. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
33. Chen, J. M., Férec, C. & Cooper, D. N. Patterns and mutational signatures of tandem base substitutions causing human inherited disease. *Hum. Mutat*. **34**, 1119–1130 (2013).
34. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* https://doi.org/10.1038/s41586-019-1907-7 (2020).
35. Haradhvala, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun*. **9**, 1746 (2018).
36. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-i. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation* (John Wiley & Sons, 2009).
37. Blei, D., Carin, L. & Dunson, D. Probabilistic topic models: a focus on graphical model design and applications to document and image analysis. *IEEE Signal Process. Mag*. **27**, 55–65 (2010).
38. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
39. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res*. **28**, 654–665 (2018).
40. Viel, A. et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
41. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol*. **242**, 10–15 (2017).
42. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
43. Saini, N. et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet*. **12**, e1006385 (2016).
44. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
45. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet*. **47**, 1067–1072 (2015).
46. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
47. Kasar, S. & Brown, J. R. Mutational landscape and underlying mutational processes in chronic lymphocytic leukemia. *Mol. Cell. Oncol*. **3**, e1157667 (2016).
48. Matsuda, T., Kawanishi, M., Yagi, T., Matsui, S. & Takebe, H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res*. **26**, 1769–1774 (1998).
49. Brash, D. E. UV signature mutations. *Photochem. Photobiol*. **91**, 15–26 (2015).
50. Hill, K. A., Wang, J., Farwell, K. D. & Sommer, S. S. Spontaneous tandem-base mutations (TBM) show dramatic tissue, age, pattern and spectrum specificity. *Mutat. Res*. **534**, 173–186 (2003).
51. Garaycoechea, J. I. et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**, 171–177 (2018).
52. Pfeifer, G. P. Formation and processing of UV photoproducts: effects of DNA sequence and chromatin environment. *Photochem. Photobiol*. **65**, 270–283 (1997).
53. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol*. **26**, 52–64 (2016).
54. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
55. Huang, M. N. et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res*. **27**, 1475–1486 (2017).
56. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
57. Olivier, M. et al. Modelling mutational landscapes of human cancers *in vitro*. *Sci. Rep*. **4**, 4482 (2014).
58. Szikriszt, B. et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol*. **17**, 99 (2016).
59. Zhivagui, M. et al. Experimental and pan-cancer genome analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans. *Genome Res*. **29**, 521–531 (2019).
60. Zámborszky, J. et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 746–755 (2017).
61. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun*. **9**, 1744 (2018).
62. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).

PCAWG Mutational Signatures Working Group

Ludmil B. Alexandrov[1], Erik N. Bergstrom[1], Arnoud Boot[4,5], Paul Boutros[25,26,27,28], Kin Chan[29], Kyle R. Covington[6,7], Akihiro Fujimoto[30], Gad Getz[2,3,21,22], Dmitry A. Gordenin[8], Nicholas J. Haradhvala[2,3], Mi Ni Huang[4,5], S. M. Ashiqul Islam[1], Marat Kazanov[31,32,33], Jaegil Kim[2], Leszek J. Klimczak[12], Nuria Lopez-Bigas[9,10,11], Michael Lawrence[2,34,35], Iñigo Martincorena[13], John R. McPherson[4,5], Sandro Morganella[13], Ville Mustonen[17,18,19], Hidewaki Nakagawa[30], Alvin Wei Tian Ng[4,5], Paz Polak[2,3,22], Stephenie Prokopec[27], Steven A. Roberts[36,37], Steven G. Rozen[4,5,23], Radhakrishnan Sabarinathan[10,14,15], Natalie Saini[8], Tatsuhiro Shibata[38,39], Yuichi Shiraishi[40], Michael R. Stratton[13], Bin Tean Teh[4,23,41,42,43], Ignacio Vázquez-García[13,44,45,46], David A. Wheeler[6,16], Yang Wu[4,5], Fouad Yousif[2] & Willie Yu[4,5]

[25]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [26]University of California Los Angeles, Los Angeles, CA, USA. [27]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [28]Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. [29]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. [30]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [31]A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. [32]Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia. [33]Skolkovo Institute of Science and Technology, Moscow, Russia. [34]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [35]Massachusetts General Hospital, Boston, MA, USA. [36]School of Molecular Biosciences, Washington State University, Pullman, WA, USA. [37]Center for Reproductive Biology, Washington State University, Pullman, WA, USA. [38]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [39]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [40]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [41]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. [42]Institute of Molecular and Cell Biology, Singapore, Singapore. [43]Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore, Singapore. [44]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [45]Department of Statistics, Columbia University, New York, NY, USA. [46]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

These online methods contain an abridged description of the methodology used in the current manuscript; extensive details about the methodology we used are provided in Supplementary Note 2. Importantly, two independently developed computational frameworks (SigProfiler and SignatureAnalyzer) based on NMF were applied separately to the examined sets of mutational catalogues. SigProfiler and SignatureAnalyzer take different approaches for deciphering mutational signatures and for assigning each signature to each sample. By using two methods, we aimed to provide a perspective on the effect that different methodologies can have on the numbers of signatures generated, signature profiles and attributions. In addition to applying SigProfiler and SignatureAnalyzer to cancer data, the tools were also applied to realistic synthetic data with known solutions.

Analysis of mutational signatures with SigProfiler

SigProfiler incorporates two distinct steps for identification of mutational signatures, based on the previously described methodology[6,11,17] (Extended Data Fig. 8). The first step (SigProfilerExtraction) encompasses a hierarchical de novo extraction of mutational signatures based on somatic mutations and their immediate sequence context, and the second step (SigProfilerAttribution) focuses on accurately estimating the number of somatic mutations associated with each extracted mutational signature in each sample. SigProfilerExtraction is an extension of a previous framework for the analysis of mutational signatures[11,17]. In brief, for a given set of mutational catalogues, the algorithm deciphers a minimal set of mutational signatures that optimally explains the proportion of each mutation type and estimates the contribution of each signature to each sample. More specifically, for each NMF iteration, SigProfilerExtraction minimizes a generalized Kullback–Leibler divergence constrained for nonnegativity (Supplementary Note 2). The algorithm uses multiple NMF iterations (in most cases 1,024) to identify the matrix of mutational signatures and the matrix of the activities of these signatures, as previously described[17]. The unknown number of signatures is determined by human assessment of the stability and accuracy of solutions for a range of values, as previously described[17]. The framework is applied hierarchically to increase its ability to find mutational signatures that generate few mutations or are present in few samples.

After signatures are discovered by SigProfilerExtraction, SigProfilerAttribution estimates their contributions to individual samples. For each examined sample, the estimation algorithm involves finding the minimum of the Frobenius norm of a constrained function using a nonlinear convex optimization programming solver using the interior-point algorithm[63]. See Supplementary Note 2 and Extended Data Fig. 8b for further details.

Analysis of mutational signatures with SignatureAnalyzer

SignatureAnalyzer uses a Bayesian variant of NMF that infers the number of signatures through the automatic relevance determination technique and delivers highly interpretable and sparse representations for both signature profiles and attributions that strike a balance between data fitting and model complexity. Further details of the actual implementation of the computational approach have previously been published[9,27,64]. SignatureAnalyzer was applied by using a two-step signature extraction strategy using 1,536 pentanucleotide contexts for SBSs, 83 indel features and 78 DBS features. In addition to the separate extraction of SBS, indel and DBS signatures, we performed a 'COMPOSITE' signature extraction based on all 1,697 features (1,536 SBS + 78 DBS + 83 indel). For SBSs, the 1,536 SBS COMPOSITE signatures are preferred; for DBSs and indels, the separately extracted signatures are preferred.

In step 1 of the two-step extraction process, global signature extraction was performed for the samples with a low mutation burden ($n = 2{,}624$). These excluded hypermutated tumours: those with putative polymerase epsilon (POLE) defects or mismatch repair defects (microsatellite instable tumours), skin tumours (which had intense UV-light mutagenesis) and one tumour with temozolomide (TMZ) exposure. Because the underlying algorithm of SignatureAnalyzer performs a stochastic search, different runs can produce different results. In step 1, we ran SignatureAnalyzer 10 times and selected the solution with the highest posterior probability. In step 2, additional signatures unique to hypermutated samples were extracted (again selecting the highest posterior probability over ten runs) while allowing all signatures found in the samples with low mutation burden, to explain some of the spectra of hypermutated samples. This approach was designed to minimize a well-known 'signature bleeding' effect or a bias of hyper- or ultramutated samples on the signature extraction. In addition, this approach provided information about which signatures are unique to the hypermutated samples, which was later used when attributing signatures to samples.

A similar strategy was used for signature attribution: we performed a separate attribution process for low- and hypermutated samples in all COMPOSITE, SBS, DBS and indel signatures. For downstream analyses, we preferred to use the COMPOSITE attributions for SBSs and the separately calculated attributions for DBSs and indels. Signature attribution in samples with a low mutation burden was performed separately in each tumour type (for example, Biliary–AdenoCA, Bladder–TCC, Bone–Osteosarc, and so on). Attribution was also performed separately in the combined microsatellite instable tumours ($n = 39$), POLE ($n = 9$), skin melanoma ($n = 107$) and TMZ-exposed samples (syn11738314). In both groups, signature availability (which signatures were active, or not) was primarily inferred through the automatic relevance determination process applied to the activity matrix H only, while fixing the signature matrix W. The attribution in samples with a low mutation burden was performed using only signatures found in the step 1 of the signature extraction. Two additional rules were applied in SBS signature attribution to enforce biological plausibility and minimize a signature bleeding: (i) allow SBS4 (smoking signature) only in lung, head and neck cases; and (ii) allow SBS11 (TMZ signature) in a single GBM sample. This was enforced by introducing a binary, signature-by-sample signature indicator matrix Z (1, allowed; 0, not allowed), which was multiplied by the H matrix in every multiplication update of H. No additional rules were applied to indel or DBS signature attributions, except that signatures found in hypermutated samples were not allowed in samples with a low mutation burden.

Application of SigProfiler and SignatureAnalyzer to synthetic data

Our goal was to evaluate SignatureAnalyzer and SigProfiler on realistic synthetic data to identify any potential limitations of these two methods. SignatureAnalyzer and SigProfiler were tested on 11 sets of synthetic data, encompassing a total of 64,400 synthetic samples, in which known signature profiles were used to generate catalogues of synthetic mutational spectra. We operationally defined 'realistic' data as those based on the characteristics of either SignatureAnalyzer's or SigProfiler's analysis of the PCAWG genome data. SignatureAnalyzer's reference signature profiles were based on COMPOSITE signatures, consisting of 1,536 types of strand-agnostic SBSs in pentanucleotide context, 78 types of DBSs and 83 types of small indels, for a total of 1,697 mutation types. SigProfiler's reference analysis was based on strand-agnostic SBSs in the context of one 5′ and one 3′ base. For each test, we generated two sets of realistic data: SigProfiler-realistic (based on SigProfiler's reference signatures and attributions) and SignatureAnalyzer-realistic (based on SignatureAnalyzer's reference signatures and attributions), as well as two other types of data that involved using SignatureAnalyzer profiles with SigProfiler attributions and vice versa.

A detailed description of each of the 11 sets of synthetic data and the results from applying SigProfiler and SignatureAnalyzer are provided in Supplementary Note 2.

Analysis of clustered mutational signatures

Somatic SBSs were considered clustered if they had intermutational distances <1,000 bp. More specifically, for each sample, an SBS mutational catalogue was generated for substitutions that were <1,000 bp from another substitution. Subsequently, the set of SBS mutational catalogues containing clustered mutations underwent de novo extraction of mutational signatures. Any novel mutational signature (one that was not previously observed in the complete SBS catalogues) was reported as a clustered mutational signature.

Better separation compared to COSMIC v.2 signatures

As described in the manuscript, all mutational signatures previously reported in COSMIC v.2 were confirmed in the new set of analyses with median cosine similarity of 0.95. However, the separation between the COSMIC v.2 mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures_v2) is much worse than the separation between the mutational signatures reported here. For example, in COSMIC v.2, signatures 5 and 16 had a cosine similarity of 0.90, making them hard to distinguish from one another. By contrast, in the current analysis, SBS5 and SBS16 have a cosine similarity of 0.65. This allows us to unambiguously assign SBS5 and SBS16 to different samples. In the current analysis, the larger number of samples has allowed the reduction of bleeding between signatures and has given more unique and easily distinguishable signatures. One can evaluate the overall separation of a set of mutational signatures by examining the distribution of cosine similarities between the signatures in the set. The signatures in COSMIC v.2 had a median cosine similarity of 0.238. By contrast, the current signatures have a much lower median cosine similarity of 0.098. This twofold reduction in similarity is highly statistically significant (*P* value 9.1×10^{-25}) and indicates a better separation between the signatures in the current analysis.

Correlations of mutational signature activity with age

Before evaluating the association between age and the activity of a mutational signature, all outliers for both age and numbers of mutations attributed to a signature in a cancer type were removed from the data. An outlier was defined as any value outside three standard deviations from the mean value. A robust linear regression model that estimated the slope of the line and whether this slope was significantly different from zero (*F* test; *P* value < 0.05) was performed using the MATLAB function robustfit (https://www.mathworks.com/help/stats/robustfit.html) with default parameters. The *P* values from the *F* tests were corrected using the Benjamini–Hochberg procedure for false discovery rates. Results are available at syn12030687 and syn20317940.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC and TCGA PCAWG Consortium are described in ref. [2], and are available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier that does not require access approval. To access information that could potentially identify participants, such as germline alleles and the underlying sequencing data, researchers will need to apply to the TCGA data access committee via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC data access compliance office (http://icgc.org/daco) for the ICGC portion of the dataset. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. For each mutational signature as extracted by SigProfiler, there is a 'vignette' that consists of plots and a short textual description at COSMIC (available at https://cancer.sanger.ac.uk/cosmic/signatures/). Beyond the core sequence data generated by the ICGC and TCGA PCAWG Consortium, other derived datasets were generated by the research reported in this paper. These derived datasets are available at Synapse (https://www.synapse.org/#!Synapse:syn11726601/wiki/513478), and are denoted by accession numbers (synXXXXXXXX). All these datasets are mirrored at https://dcc.icgc.org/releases/PCAWG/mutational_signatures/ with full links, filenames, accession numbers and descriptions as detailed in Supplementary Table 1. These datasets include (1) CSV files comprising all catalogues of observed mutational spectra that were used as input to signature extraction (syn11801889), (2) CSV files and plots of signatures extracted by SigProfiler (syn11738306) and SignatureAnalyzer (syn11738307), (3) CSV files with estimates of the numbers of mutations generated by each signature in individual tumours (syn11804065), (4) estimates of the probability that each signature was responsible for each mutational type (for example, CTG>CAG) in individual tumours (syn11804068) and (5) synthetic test input data plus the results of tests of signature extraction (discovery) on the synthetic test data (syn18497223). All derived datasets are open access, and can be downloaded without registration or logging in.

Code availability

SigProfiler is available both as a MATLAB framework and as a Python package. In both cases, SigProfiler is a fully functional, free and open-source tool distributed under the permissive 2-Clause BSD License. SigProfiler in MATLAB can be downloaded from: https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler. SigProfiler in Python can be downloaded from: https://github.com/AlexandrovLab/SigProfilerExtractor. SignatureAnalyzer code is available at https://github.com/broadinstitute/getzlab-SignatureAnalyzer (github.com). The code used to generate the synthetic data and summarize SignatureAnalyzer and SigProfiler results is open source and freely available as the SynSig package: https://github.com/steverozen/SynSig/tree/v0.2.0 under the GNU General Public License v.3.0. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v.3.0, which allows for reuse and distribution.

63. Byrd, R. H., Hribar, M. E. & Nocedal, J. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optim.* **9**, 877–900 (1999).
64. Tan, V. Y. & Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).

Acknowledgements The results here are based in part on data generated by the TCGA research network (http://cancergenome.nih.gov/) and the ICGC and TCGA PCAWG network. This work was supported by Wellcome grant reference 206194 (M.R.S.), Singapore National Medical Research Council grants NMRC/CIRG/1422/2015 and MOH-000032/MOH-CIRG18may-0004 and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes (M.N.H., A.W.T.N., Y.W., A.B. and S.G.R.), US National Institute of Health Intramural Research Program Project Z1AES103266 (D.A.G.), the European Research Council Consolidator Grant 682398 (N.L.-B.), US National Cancer Institute U24CA143843 (D.A.W.) and Cancer Research UK Grand Challenge Award C98/A24032 (E.N.B., S.M.A.I., L.B.A. and M.R.S.). G.G and J.K were partially supported by the National Cancer Institute grants U24CA210999 and U24CA143845. G.G. was partially supported by the Paul C. Zamecnick, MD, Chair in Oncology at the Massachusetts General Hospital Cancer Center. N.J.H. and G.G. were partially supported by G.G.'s funds at the Broad Institute and Massachusetts General Hospital. N.J.H. was partially funded by the Molecular Biophysics Training Grant NIH/ NIGMS T32 GM008313 (PI: Venkatesh N. Murthy).We acknowledge the contributions of the many clinical networks across the ICGC

and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects. The members of the PCAWG Consortium are listed in Supplementary Note 1.

Author contributions The ICGC and TCGA contributed collectively to this work under the guidance of PCAWG Steering and Executive Committees, and the Ethics and Legal Working Group. The International Cancer Genome Consortium and TCGA tumour specific providers provided tumour and matched non-tumour samples, and the PCAWG Technical Working Group, the PCAWG Quality Control Working Group and the PCAWG Novel Somatic Mutation Calling Methods Working Group provided standardized mutation calls for the 2,780 PCAWG whole genomes. G.G., S.G.R. and M.R.S. were project leaders; L.B.A., G.G., S.G.R. and M.R.S. obtained funding for this study; L.B.A., J.K., N.J.H., G.G., S.G.R. and M.R.S. designed this study; M.N.H., A.W.T.N., A.B., E.N.B., J.R.M. and S.G.R. collected and prepared data for analysis; L.B.A., J.K., E.N.B. and S.M.A.I. created mutational signature analysis software; L.B.A., J.K., N.J.H., A.W.T.N., A.B., K.R.C., D.A.G., N.L.-B., L.J.K., S.M., R.S., D.A.W., V.M., G.G., S.G.R. and M.R.S. analysed data and reviewed results; L.B.A., J.K., N.J.H., G.G., S.G.R. and M.R.S. wrote the paper; L.B.A., J.K., N.J.H., M.N.H. and A.W.T.N. created figures; and Y.W. and S.G.R. generated synthetic data and benchmarked signature analysis software.

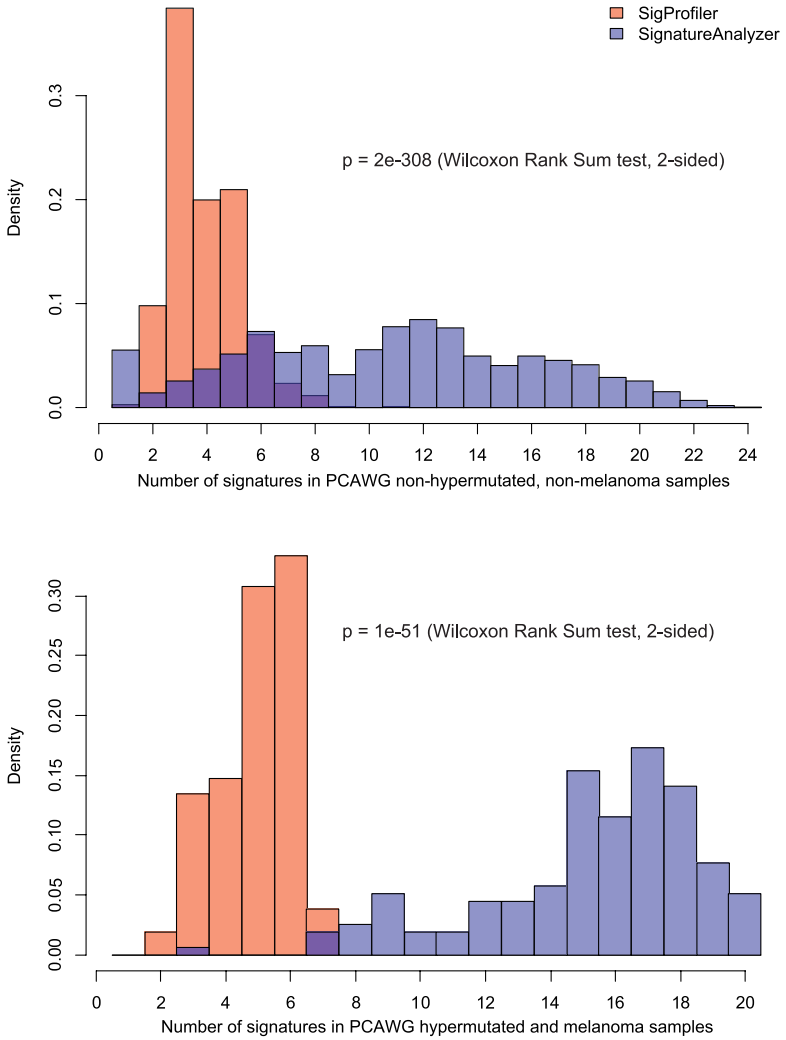
Competing interests G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect and POLYSOLVER. All the other authors have no competing interests.

Additional information

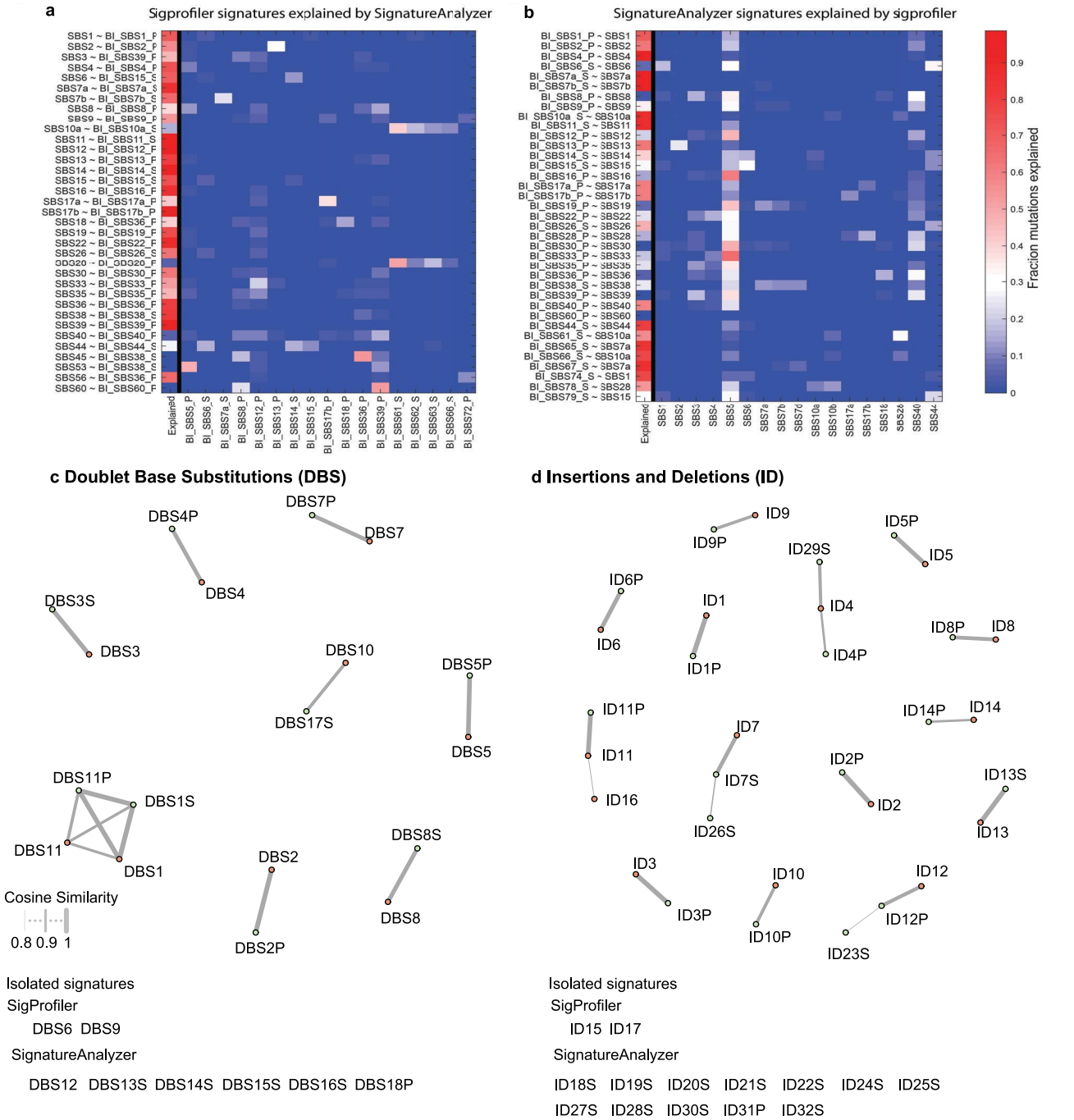
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1943-3.

Correspondence and requests for materials should be addressed to S.G.R. or M.R.S.

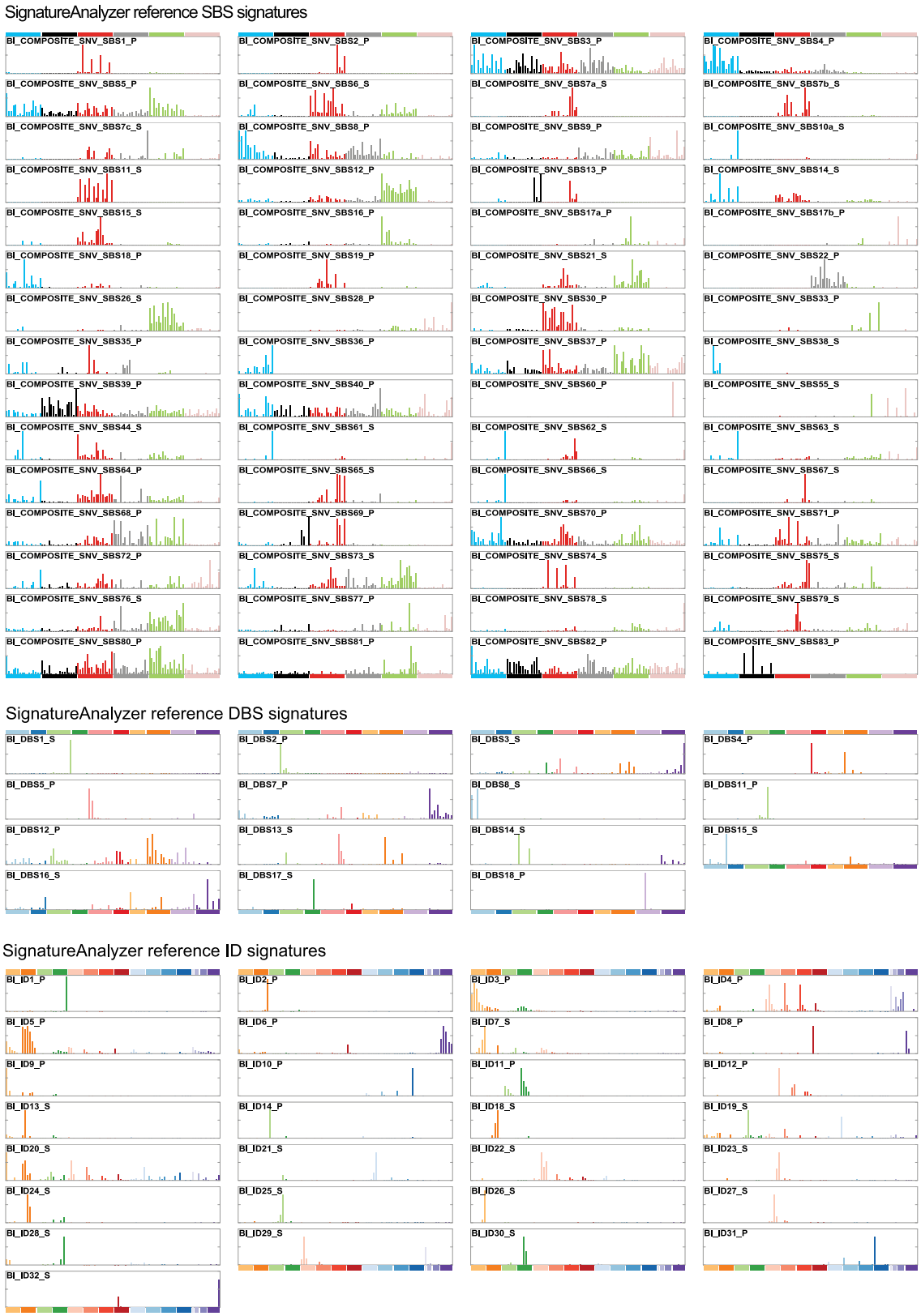
Peer review information *Nature* thanks Arul Chinnaiyan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



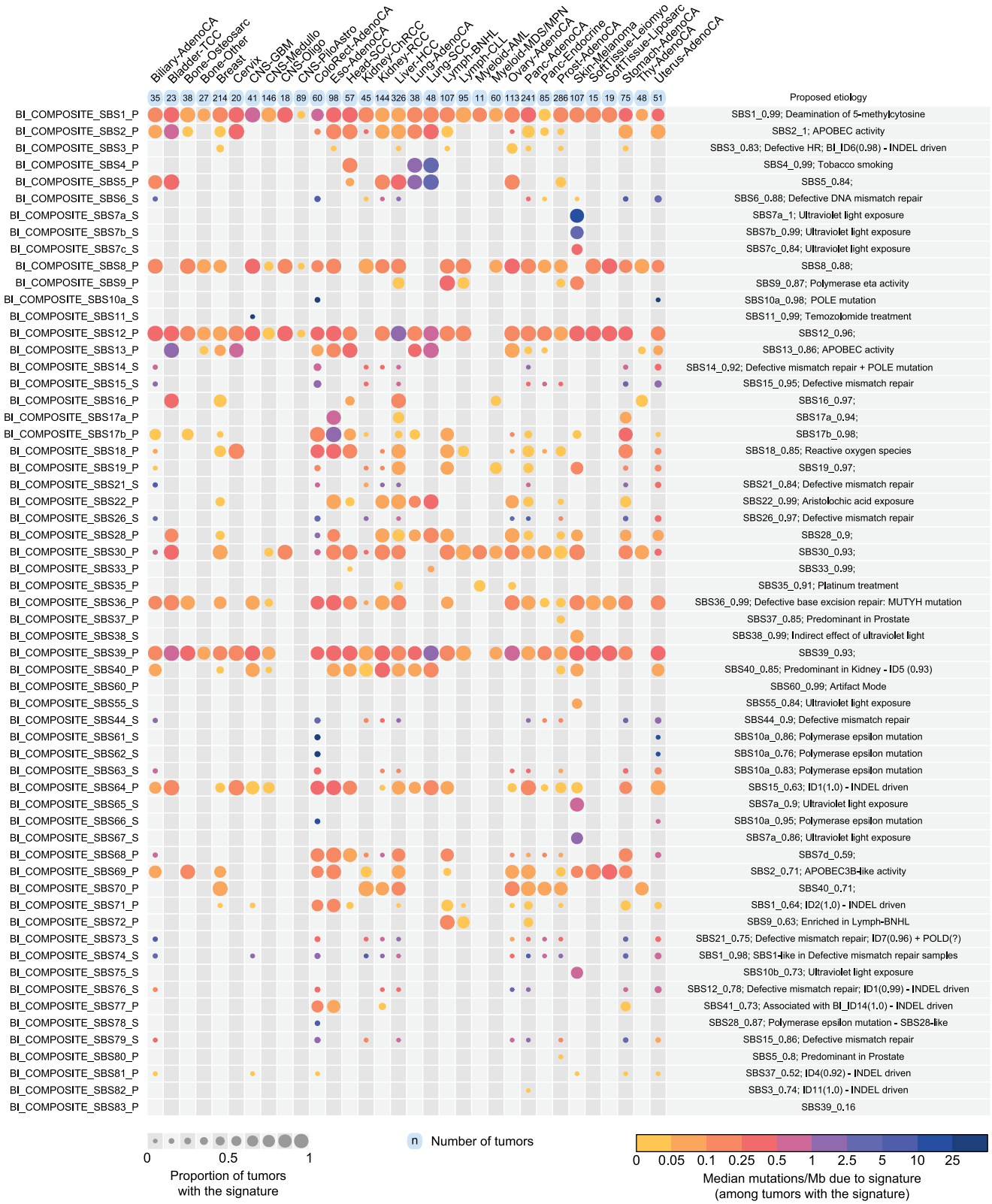
Extended Data Fig. 1 | Histogram of the number of signatures attributed in each of 2,780 PCAWG samples by SigProfiler and SignatureAnalyzer. Hypermutated tumours and melanomas (156) are listed at syn11738314.



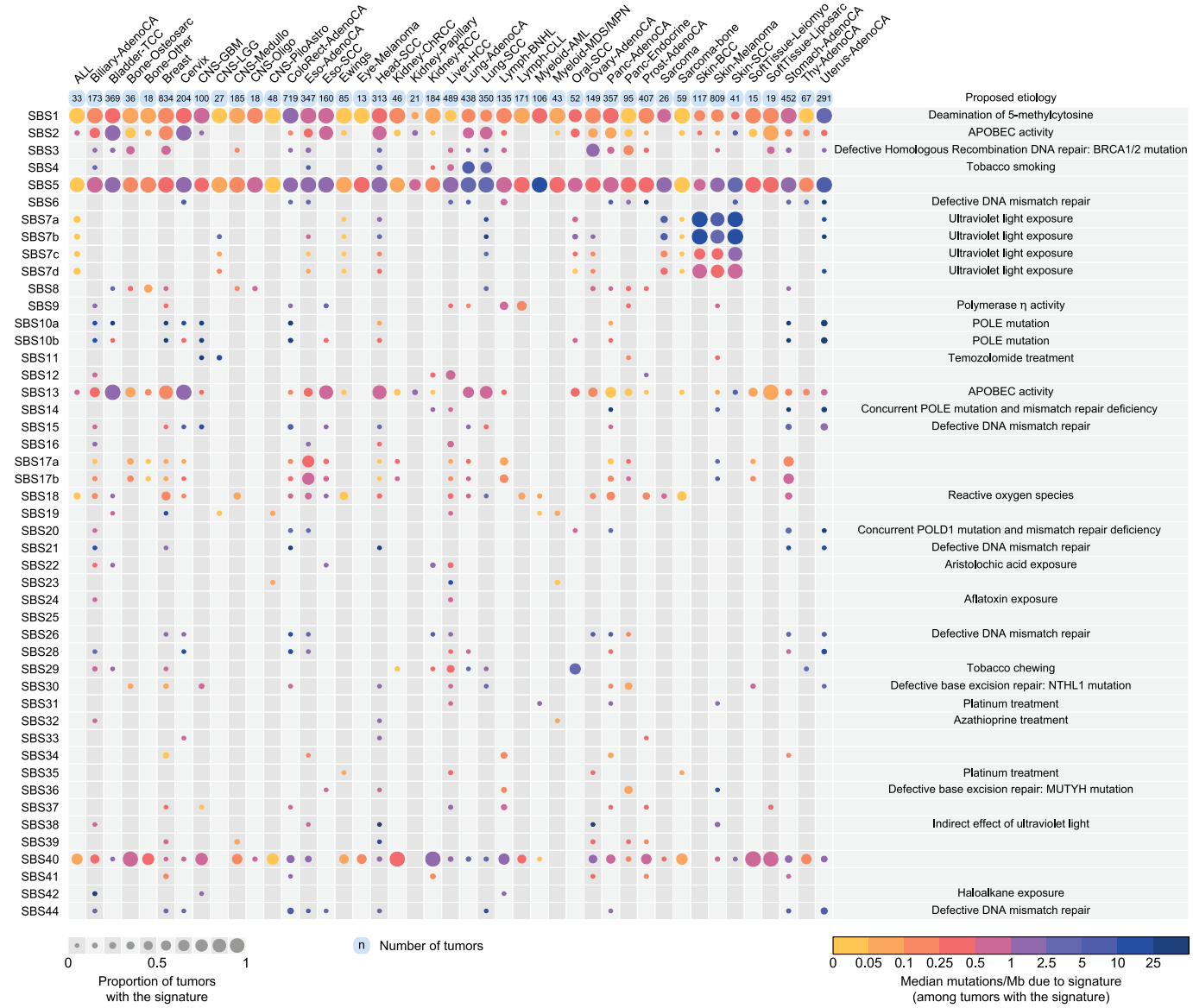
Extended Data Fig. 2 | Comparisons between results of SigProfiler and SignatureAnalyzer. a, **b**, Comparison of the attributions for corresponding SigProfiler (**a**) and SignatureAnalyzer (**b**) signatures. Each one of the SBS signatures extracted by SigProfiler and SignatureAnalyzer was paired with the signature of highest cosine similarity in the extraction by the other method (if one with >0.85 cosine similarity exists). The first column of the plot corresponds to the fraction of mutations assigned by one method (summed across samples and mutation types) that was also assigned by the other method. The remaining mutations were then redistributed to the other signatures in the extraction, weighted by their relative probabilities of having been generated by each signature and the resulting fraction of mutations was then plotted. Signatures on the *x* axis are shown only if they contribute at least a 0.1 fraction of mutations to at least one signature on the *y* axis. **c**, **d**, Cosine similarities between SigProfiler and SignatureAnalyzer DBS (**c**) and indel (**d**) signatures. Brown nodes represent SigProfiler signatures; green nodes represent SignatureAnalyzer signatures. Matches with cosine similarities > 0.8 are shown as edges; the width of the edge indicates the strength of the similarity. The locations of the nodes have no meaning. Signatures with no matches of >0.8 cosine similarity are shown below. SigProfiler ID15 and ID17 were extracted from data that were not analysed by SignatureAnalyzer. The suffix 'P' on a SignatureAnalyzer signature name indicates a signature extracted from non-hypermutated, non-melanoma tumours. The suffix 'S' on a SignatureAnalyzer signature name indicates a signature extracted from hypermutated or melanoma tumours.



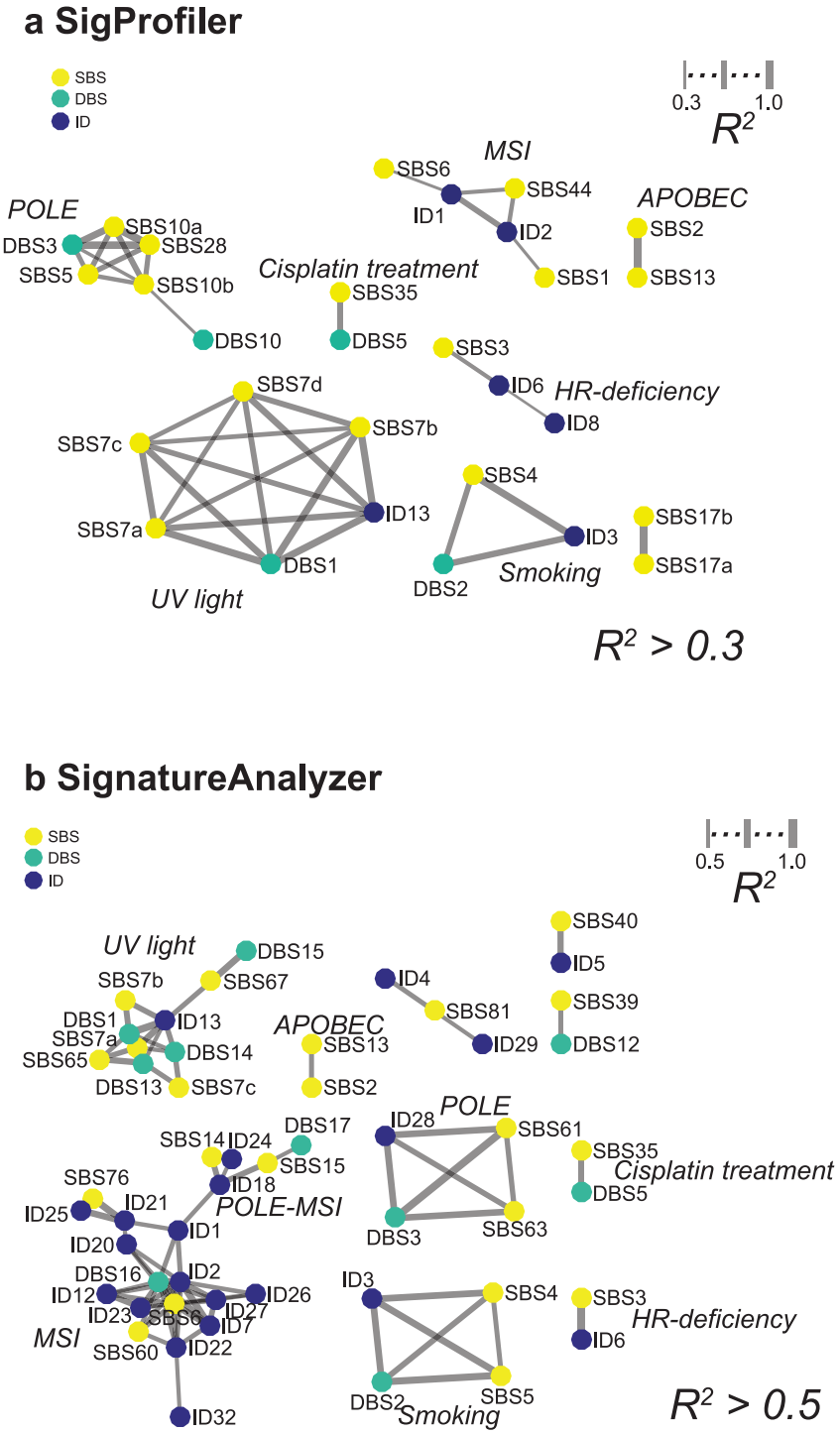
Extended Data Fig. 3 | SignatureAnalyzer reference signatures. The classifications of each mutation type (SBS, 96 classes; DBS, 78 classes; and indels, 83 classes) are described in the main text.



Extended Data Fig. 4 | The number of SBS mutations attributed to each mutational signature for each cancer type over the PCAWG tumours by SignatureAnalyzer. Conventions are as in Fig. 3; see this figure for explanation.

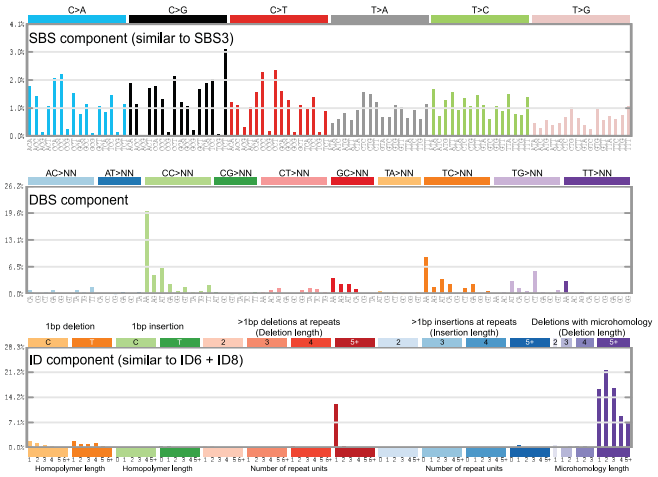


Extended Data Fig. 5 | The number of SBS mutations attributed to each mutational signature to each cancer type over the complete set of PCAWG and non-PCAWG cancer samples analysed by SigProfiler. Conventions are as in Fig. 3; see this figure for explanation.



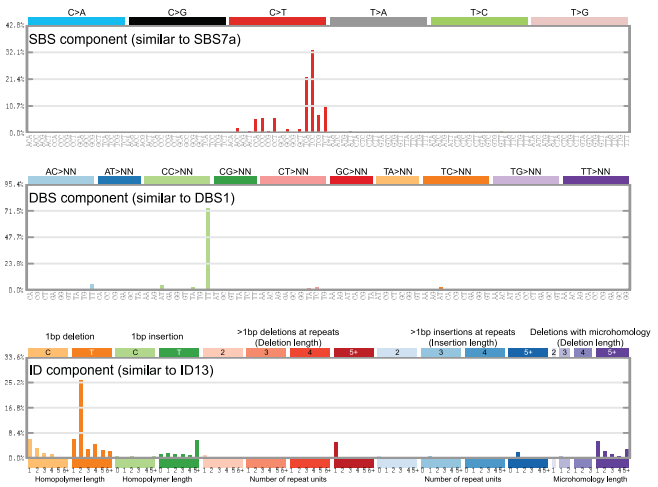
Extended Data Fig. 6 | Associations between SBS, DBS and indel signature activities for SigProfiler and SignatureAnalyzer. a, **b**, Each node represents an SBS (light green), DBS (dark green) or indel (black) signature. Any two signatures with sample attributions that significantly correlated with $R^2 > 0.3$ (SigProfiler) (**a**) or > 0.5 (SignatureAnalyzer) (**b**) are connected by edges. Edge widths are proportional to the strength of the correlation. Signatures with no significant correlation to any other signature above the relevant threshold are not shown. Signature locations are fit for display purposes only, and do not indicate similarity.

Composite-3

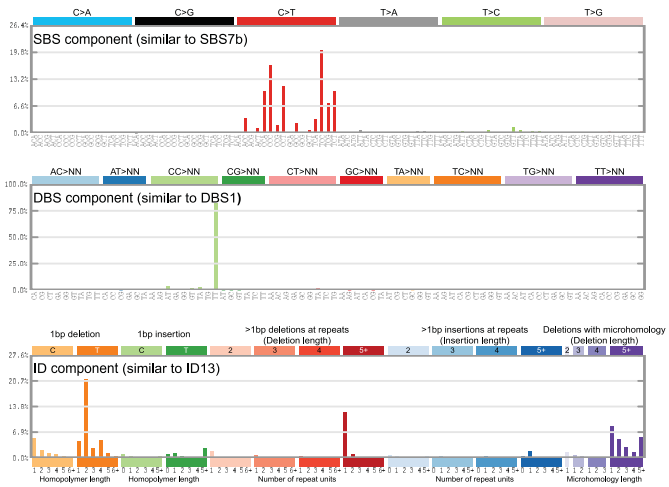


Composite-4

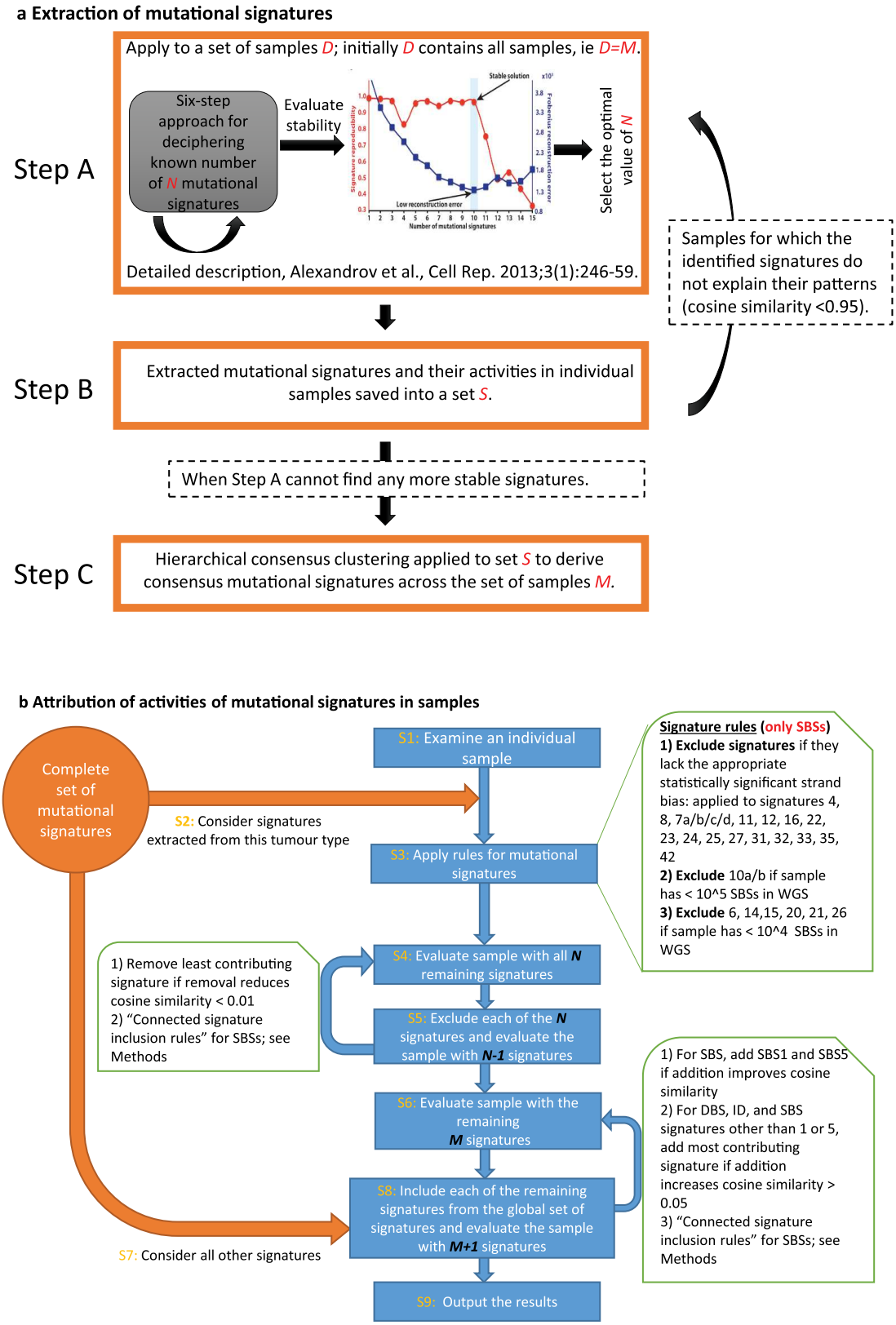
Composite-7a



Composite-7b



Extended Data Fig. 7 | Mutational signatures extracted from the COMPOSITE feature set consisting of the concatenation of SBSs in pentanucleotide context, DBSs and indels. For each of the 4 COMPOSITE mutational signatures shown, the top panel shows the SBS signature in pentanucleotide context (1,536 mutation classes) after being collapsed to 96 SBS mutation classes, the middle panel is the co-extracted DBS signature and the bottom panel is the co-extracted indel signature. There are similarities between the DBS portion of Composite-4 and DBS2, and between the indel portion of Composite-4 and ID3; other similarities are noted in the figure.



Extended Data Fig. 8 | SigProfiler signature extraction and attribution. A full description is provided in Supplementary Note 2. **a**, Procedure for extracting (discovering) mutational signatures. Step A, apply the approach to a set of samples *D*; initially *D* contains all samples (that is, *D* = *M*). This step has previously been described in detail[17]. Step B, solution evaluation and re-iteration. Extracted mutational signatures and their activities in individual samples are saved into a set (*S*). The activity of any signature that does not increase the cosine similarity of a sample by > 0.01 was removed from the sample (assigned a value of 0). Step A is repeated for all samples for which the identified signatures do not explain their patterns (cosine similarity < 0.95). The algorithm continues to step C when step A cannot find any stable signatures. Step C, clustering of mutational signatures. Hierarchical consensus clustering was applied to the set *S* to derive the consensus mutational signatures across the set of samples *M*. **b**, Attribution of activities of mutational signatures in samples.

Extended Data Table 1 | The number of DBSs is proportional to the number of SBSs, with few exceptions

Covariate (including cancer type)	Coefficient estimate	Coefficient std. error	t value	Pr(> |t|)	Tumour count
(Intercept)	5.60E+00	8.80E+01	0.1	0.9	NA
SBS.count	3.70E-03	1.30E-04	29.8	<2e-16	NA
Biliary-AdenoCA	(reference)				35
Bladder-TCC	1.30E+01	1.40E+02	0.1	0.9	23
Bone-Benign	-6.30E+00	1.60E+02	0	1	16
Bone-Epith	2.20E+00	1.80E+02	0	1	11
Bone-Osteosarc	2.20E+00	1.20E+02	0	1	38
Breast-AdenoCA	6.20E+00	9.50E+01	0.1	0.9	198
Breast-DCIS	4.20E+00	3.10E+02	0	1	3
Breast-LobularCA	-8.20E+00	1.70E+02	0	1	13
Cervix-AdenoCA	-7.90E+00	3.80E+02	0	1	2
Cervix-SCC	-1.10E+01	1.50E+02	-0.1	0.9	18
CNS-GBM	-2.80E+01	1.20E+02	-0.2	0.8	41
CNS-Medullo	-7.00E+00	9.80E+01	-0.1	0.9	146
CNS-Oligo	-1.00E+01	1.50E+02	-0.1	0.9	18
CNS-PiloAstro	-5.90E+00	1.00E+02	-0.1	1	89
ColoRect-AdenoCA	-4.10E+02	1.10E+02	-3.7	3.00E-04	60
Eso-AdenoCA	-1.60E+01	1.00E+02	-0.2	0.9	98
Head-SCC	5.30E+01	1.10E+02	0.5	0.6	57
Kidney-ChRCC	-3.10E+00	1.20E+02	0	1	45
Kidney-RCC	5.60E+01	9.80E+01	0.6	0.6	144
Liver-HCC	7.80E+01	9.20E+01	0.8	0.4	326
Lung-AdenoCA	5.00E+02	1.20E+02	4.1	4.00E-05	38
Lung-SCC	5.80E+02	1.20E+02	5.1	4.00E-07	48
Lymph-BNHL	1.00E+01	1.00E+02	0.1	0.9	107
Lymph-CLL	-4.30E+00	1.00E+02	0	1	95
Myeloid-AML	-1.90E+00	1.80E+02	0	1	11
Myeloid-MDS	-8.00E+00	2.70E+02	0	1	4
Myeloid-MPN	-7.40E+00	1.10E+02	-0.1	0.9	56
Ovary-AdenoCA	3.60E+01	1.00E+02	0.4	0.7	113
Panc-AdenoCA	-8.30E-01	9.40E+01	0	1	241
Panc-Endocrine	-5.70E+00	1.00E+02	-0.1	1	85
Prost-AdenoCA	2.50E+00	9.30E+01	0	1	286
Skin-Melanoma	1.70E+03	1.00E+02	16.5	<2e-16	107
SoftTissue-Leiomyo	6.00E+00	1.60E+02	0	1	15
SoftTissue-Liposarc	7.80E+00	1.50E+02	0.1	1	19
Stomach-AdenoCA	-3.00E+01	1.10E+02	-0.3	0.8	75
Thy-AdenoCA	-4.80E+00	1.20E+02	0	1	48
Uterus-AdenoCA	-1.20E+02	1.10E+02	-1.1	0.3	51

The exceptions are colorectal adenocarcinoma (Colorect–AdenoCA), lung adenocarcinoma (Lung–AdenoCA), lung squamous cell carcinoma (Lung–SCC) and skin–melanoma, as analysed by the following linear regression (computed by an R function call): glm(DBS.count ~ SBS.count + Cancer.Type). This function call fits a model in which the number of DBSs depends linearly on the number of SBSs and on the cancer type. *P* values associated with the coefficients are two-sided.

Extended Data Table 2 | Numbers of insertion and deletion mutations due to ID1, ID2 and all other indel signatures in hypermuted and non-hypermutated tumours

	Hypermutators		Non-hypermutators		All Tumours	
Signature	**Count**	**Fraction**	**Count**	**Fraction**	**Count**	**Fraction**
ID1	593,935	0.236	399,633	0.276	993,568	0.250
ID2	1,838,867	0.730	252,893	0.174	2,091,760	0.527
ID1+ID2	2,432,802	0.966	652,526	0.450	3,085,328	0.777
Other ID signatures	85,038	0.034	797,964	0.550	883,002	0.223
Total	2,517,840	1	1,450,490	1	3,968,330	1

nature research

Corresponding author(s): Steven G. Rozen and Michael R. Stratton

Last updated by author(s): Oct 8, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data in this study were those reported in https://www.biorxiv.org/content/early/2017/07/12/162784.full.pdf+html (the PCAWG marker paper) and in the publications cited at https://www.synapse.org/#!Synapse:syn11801788. For the larger PCAWG Consortium, data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at https://www.overture.bio. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to contact@overture.bio.
Data analysis	SigProfiler is available both as a MATLAB framework and as a Python package. In both cases, SigProfiler is fully functional, free, and open-source tool distributed under the permissive 2-Clause BSD License. SigProfiler in MATLAB can be downloaded from: https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler SigProfiler in Python can be downloaded from: https://github.com/AlexandrovLab/SigProfilerExtractor. SignatureAnalyzer code is available at https://www.synapse.org/#!Synapse:syn11801492. The code used to generate the synthetic data and summarize SignatureAnalyzer and SigProfiler results is open-source and freely available as the SynSig package: https://github.com/steverozen/ SynSig/tree/v0.2.0 under the GPL3 license. For the larger PCAWG Consortium, the workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACEseq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Derived data are available at https://www.synapse.org/#!Synapse:syn11726601/wiki/513478. All figures and extended data figures have associated raw data at that site.

For the larger PCAWG Consortium, WGS somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	From a statistical perspective this was an exploratory study, and there were no pre-defined hypothesis tests for which sample-size power calculations would have been appropriate. The sample size was determined by numbers of tumour genomes and exomes represented by publicly available somatic mutation data. These data consisted of the ICGC Pan Cancer whole genome mutation data, the TCGA MC3 whole exome mutation data, and additional mutation data as described in https://www.synapse.org/#!Synapse:syn11801788. This was an unsupervised analysis, and therefore we extracted as many signatures as possible from all the available data. This enabled a substantial increment over previously available sets of mutational signatures, especially with respect to double base substitution (DBS) signatures and insertion/deletion (ID) signatures. For the larger PCAWG Consortium, the Consortium compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. We collected genome data from 2,834 donors, representing all ICGC and TCGA donors that met these criteria at the time of the final data freeze in autumn 2014.
Data exclusions	From a statistical perspective this was an exploratory study, and there were no pre-defined hypothesis tests for which pre-defined data exclusion criteria would have been appropriate. Therefore, no data were excluded from analysis by our algorithms. For the larger PCAWG Consortium, after quality assurance, data from 176 donors were excluded as unusable. Reasons for data exclusions included inadequate coverage, extreme bias in coverage across the genome, evidence for contamination in samples and excessive sequencing errors (for example, through 8-oxoguanine).
Replication	This was not an experimental study, and there were no experimental replicates. For the larger PCAWG Consortium, in order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, we performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.
Randomization	There were no experimental groups in this study; the question of allocation to experimental groups is not applicable. For the larger PCAWG Consortium, no randomisation was performed.

Blinding	There was no allocation to experimental groups; the question of whether investigators were blinded to allocation is not applicable. For larger PCAWG Consortium, no blinding was undertaken.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	For the PCAWG Consortium data, patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced. The non-PCAWG analyses used previously published data.
Recruitment	For the PCAWG Consortium data, patients were recruited by the participating centres following local protocols.
Ethics oversight	For the PCAWG Consortium data, the Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) — *Specify: functional, structural, diffusion, perfusion.*

Field strength — *Specify in Tesla*

Sequence & imaging parameters — *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition — *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software — *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization — *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template — *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal — *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring — *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings — *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested — *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) — *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction — *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity — *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis — *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis — *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*