Mutational signatures: emerging concepts, caveats and clinical applications

Gene Koh[1,2], *Andrea Degasperi*[1,2], *Xueqing Zou*[1,2], *Sophie Momen*[1,2] *and Serena Nik-Zainal*[1,2]✉

Abstract | Whole-genome sequencing has brought the cancer genomics community into new territory. Thanks to the sheer power provided by the thousands of mutations present in each patient's cancer, we have been able to discern generic patterns of mutations, termed 'mutational signatures', that arise during tumorigenesis. These mutational signatures provide new insights into the causes of individual cancers, revealing both endogenous and exogenous factors that have influenced cancer development. This Review brings readers up to date in a field that is expanding in computational, experimental and clinical directions. We focus on recent conceptual advances, underscoring some of the caveats associated with using the mutational signature frameworks and highlighting the latest experimental insights. We conclude by bringing attention to areas that are likely to see advancements in clinical applications.

The concept of mutational signatures was introduced in 2012 following the demonstration that analysis of all substitution mutations in a set of 21 whole-genome-sequenced (WGS) breast cancers could reveal consistent patterns of mutagenesis across tumours[1] (FIG. 1a). These patterns were the physiological imprints of DNA damage and repair processes that had occurred during tumorigenesis and could distinguish *BRCA1*-null and *BRCA2*-null tumours from sporadic breast cancers. Subsequently, a landmark study applied this principle on ~500 WGS and ~6,500 whole-exome-sequenced tumours across 30 cancer types and revealed 21 distinct single-base substitution mutational signatures (SBSs)[2]. Recently, an updated analysis of ~4,600 WGS and ~19,000 whole-exome-sequenced samples raised the number of known SBSs to 49 (REF.[3]). Further complexity, including possible tissue specificities for some mutational signatures, has also been demonstrated[4].

Today, mutational signature analyses have become a standard component of genomic studies because they can reveal environmental and endogenous sources of mutagenesis in each tumour. Indeed, this nascent field is gaining prominence and heading towards being used in a clinically meaningful way.

While these are positive trends, it is appropriate to ask whether there are limitations to this substantially broadening field. As an increasing number of signatures of different mutation classes are being reported[3–6] (FIG. 1), correlations are being drawn between them and various factors, such as age and exposures to acid reflux or drug therapies, in an attempt to decipher causes[7–9]. However, the origins of many signatures remain cryptic. Furthermore, while earlier analyses reported single signatures, for example of UV radiation (that is, SBS7)[2], more recent studies reported multiple versions of these signatures (that is, SBS7a, SBS7b, SBS7c and SBS7d)[3], leading the community to question whether some findings are reflective of biology or are simply mathematical artefacts. Efforts to experimentally validate these abstract mathematical results are therefore warranted. Non-expert users of mutational signatures require insights into practical issues and caveats in the use of signature analysis frameworks. For clinicians, using mutational signatures reliably for clinical stratification is critical.

This Review appraises topical developments in the field of mutational signatures. We do not address the pros and cons of various mathematical models, which have been reviewed elsewhere[10,11]; rather, we seek to emphasize biological principles, highlight recent experimental work and discuss developments towards clinical applications.

Mutational signatures: current knowledge

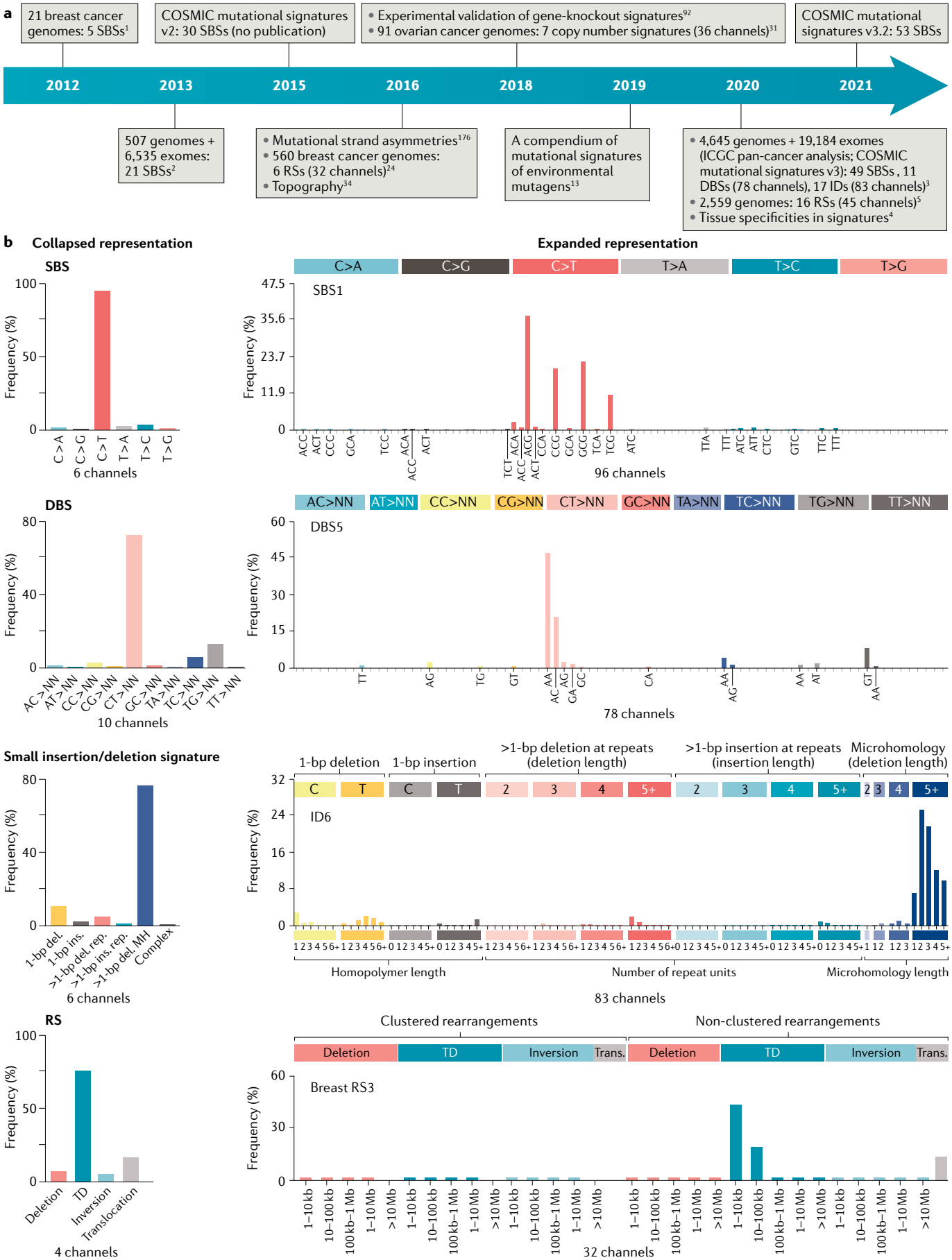
In this section, we bring readers up to date with mutational signatures, focusing on signatures of different classes. In trying to identify mutational signatures in a data set, a 'global' approach can be adopted, where signatures from all cancers, irrespective of the tissue type, are aggregated and averaged to derive a set of consensus

[1]*Department of Medical Genetics, School of Clinical Medicine, University of Cambridge, Cambridge, UK.*

[2]*MRC Cancer Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK.*

✉*e-mail: sn206@cam.ac.uk*

https://doi.org/10.1038/s41568-021-00377-7


a
21 breast cancer genomes: 5 SBSs[1]
COSMIC mutational signatures v2: 30 SBSs (no publication)
• Experimental validation of gene-knockout signatures[92]
• 91 ovarian cancer genomes: 7 copy number signatures (36 channels)[31]
COSMIC mutational signatures v3.2: 53 SBSs
2012
2013
2015
2016
2018
2019
2020
2021
507 genomes + 6,535 exomes: 21 SBSs[2]
• Mutational strand asymmetries[176]
• 560 breast cancer genomes: 6 RSs (32 channels)[24]
• Topography[34]
A compendium of mutational signatures of environmental mutagens[13]
• 4,645 genomes + 19,184 exomes (ICGC pan-cancer analysis; COSMIC mutational signatures v3): 49 SBSs , 11 DBSs (78 channels), 17 IDs (83 channels)[3]
• 2,559 genomes: 16 RSs (45 channels)[5]
• Tissue specificities in signatures[4]
b
Collapsed representation
Expanded representation
SBS
Frequency (%)
6 channels
C>A C>G C>T T>A T>C T>G
SBS1
96 channels
DBS
10 channels
AC>NN AT>NN CC>NN CG>NN CT>NN GC>NN TA>NN TC>NN TG>NN TT>NN
DBS5
78 channels
Small insertion/deletion signature
1-bp del. 1-bp ins. >1-bp del. rep. >1-bp ins. rep. >1-bp del. MH Complex
6 channels
1-bp deletion
1-bp insertion
>1-bp deletion at repeats (deletion length)
>1-bp insertion at repeats (insertion length)
Microhomology (deletion length)
ID6
Homopolymer length
Number of repeat units
Microhomology length
83 channels
RS
Deletion TD Inversion Translocation
4 channels
Clustered rearrangements
Non-clustered rearrangements
Deletion TD Inversion Trans.
Breast RS3
1–10 kb 10–100 kb 100 kb–1 Mb 1–10 Mb >10 Mb
32 channels

◀ Fig. 1 | **Conceptual developments and visualization of mutational signatures.** **a** | Chronological account of how concepts in the field of mutational signatures have evolved over time[1–5,13,24,31,34,92,176]. Key proof-of-concept experimental studies focusing on human samples are also shown. **b** | With adequate power, the preferred method for presenting single-base substitution mutational signatures (SBSs; for example, SBS1) is via the 96-channel method (see Supplementary Fig. 1 for fully labelled channels). It is also possible to expand the method to 1,536 channels (not shown) or to reduce it to only six channels. Double-base substitution signatures (DBSs; for example, DBS5) can be defined by 78 strand-agnostic channels, or when the burden is low, by ten duplet motifs. Small insertion or deletion (indel) (less than 100 bp) signatures (for example, indel signature 6 (ID6)) are broadly classified by type (that is, insertion, deletion or complex) and — when of a single base — as C or T, and according to the length of the mononucleotide repeat tract where they occur. Longer indels are classified by whether they occur at repeats or have microhomology at indel junctions. With enough power, the motif sizes and nucleotides affected can also be considered. Rearrangement signatures (RSs; for example, breast RS3) can be categorized on the basis of the four types of rearrangements and how they are regionally clustered and with further consideration of the size of the rearranged fragment. Collapsed presentations may be necessary where the mutation burden is low (for example, in experimental systems). del., deletion; ins., insertion; MH, microhomology; rep., repeat; TD, tandem duplication; trans., translocation. Part **b** adapted with permission from REF.[3], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/), and REF.[24], Springer Nature Limited.

signatures[2,3]. This approach presupposes that more samples provide more power for discerning new signatures. However, an aggregated analysis also assumes that signatures are identical across all tissues, ignoring possible tissue-specific signature properties that reflect organ-specific biology, highlighted as probable recently[4]. Indeed, the number of samples per tumour type has been imbalanced in past analyses, resulting in signatures of certain tissue types being more influential and thereby introducing potential bias[2,3]. By contrast, a 'local' approach restricts signature extractions within individual tissue types and subsequently compares locally extracted signatures between different organs[4]. This permits natural variation among different tissues to emerge. Here, we use the terms 'global' and 'local' when discussing signatures.

Another important feature is the 'channels' that classify mutations within substitution signatures, insertion or deletion (indel) signatures (IDs) and rearrangement signatures (RSs). Historically, single-base substitutions were classified by incorporating the flanking sequence contexts of each possible substitution, resulting in a 96-channel pattern for SBSs[1,12] (FIG. 1b). Double-base substitution signatures (DBSs) are defined by 78 channels[3,13]. Indel and rearrangement channel classifications are described in more detail herein, given that they are newer and not yet widely adopted.

Base substitution signatures. There is an ever-growing list of reported SBSs and DBSs (FIG. 1a). There is also an increasing number of inference techniques for identifying them[4,14–19]. Regardless of the algorithms used for signature identification, common signatures tend to be consistently identifiable in most cohorts examined, for example SBS1 (REF.[2]), caused by deamination of 5-methylcytosine. Likewise, signatures associated with environmental exposures tend to be immediately demonstrable (for example, UV-associated SBS7 and DBS1)[3]. Deficiencies in specific DNA repair pathways produce marked mutagenesis such as mismatch repair (MMR) deficiency-associated SBS26 and SBS44 (REF.[3]), and some endogenous signatures are highly distinctive and easily discernible, such as the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC)-associated SBS2 and SBS13 (REFS[2,3,20]). Rare mutational processes that are present at low population frequencies may be more challenging to extract and will reveal themselves only if they are present in the specific cohort examined (for example, treatment-associated signatures).

Our purpose in this Review is to focus on guiding principles and not to discuss all signatures individually. This information is accessible from various online resources, such as COSMIC or Signal, where exact contents may change over time. A reference set that is often used in analyses is the COSMIC v2 collection of 30 signatures. The COSMIC v3.1 collection, released in 2020, brought the total number of signatures to 49 (REF.[3]). Several signatures were noted as treatment related (for example, SBS31 and SBS35, associated with platinum; SBS90 attributed to duocarmycin), and some signatures were noticeably modified in the COSMIC v3.1 collection relative to the COSMIC v2 collection (for example, SBS1 and SBS16). Whether these amendments are true biologically or are merely a mathematical outcome is less clear and awaits independent verification. Importantly, with an increasing number of reference signatures come difficulties with accurate use and interpretation of these (discussed later).

Indel signatures. Compared with substitutions, small indels (less than 100 bp) are underexplored because of historical difficulty in obtaining high-quality indel data. Nevertheless, indels are common in cancers, occurring at ~10% of the frequency at which substitutions occur, and their genomic locations and sequence compositions are non-random[1,2]. Therefore, indels can also be presented as biologically insightful signatures. Indels cannot always be pinned to a defined coordinate with the same precision as substitutions because it is impossible to pinpoint the deleted or inserted position in a polynucleotide repeat tract. Thus, indels have been classified more simply, on the basis of their type (deletion, insertion or complex), size and whether there are features at indel junctions that could reveal biological underpinnings[1]. For example, 1-bp indels occurring at repetitive tracts commonly arise from strand slippage during replication, whereas indels that share a microhomologous sequence with the flanking sequence are thought to be scars of imperfect repair of double-strand breaks by alternative end-joining processes[21]. This simple classification formed the basis for identification of cancers with MMR deficiency and homologous recombination deficiency (HRD)[1,22].

To extract IDs, a global analysis of 2,780 cancers of multiple tissue types was performed on indels classified according to a set of 83 channels[3], which were expansions of the previous indel classification[1,2]; for example, single-base indels were classified by the numerical length of the repetitive tract in which they occurred (FIG. 1b). Seventeen IDs were reported[3]. ID6 represented a microhomology-mediated deletion signature seen in *BRCA1*-mutated and *BRCA2*-mutated cancers described previously[1]. ID1, ID2 and ID7 were

Homologous recombination deficiency
(HRD). An inability to repair DNA double-strand breaks via homologous recombination-based repair mechanisms, typically caused by germ line or somatic BRCA gene mutations.

repeat-mediated IDs highly elevated in tumours with mutations in the proofreading domains of *POLE* or *POLD1* and/or MMR deficiency. Caused by replication slippage, ID1 and ID2 A or T indels at long poly(dA:dT) tracts were also found in most samples, including normal tissues[23]. ID3 was associated with tobacco smoking, and ID13 was associated with UV exposure. ID8 was believed to be the footprint of non-homologous end joining based on 1-bp microhomology or the absence of microhomology at indel junctions[3]. A subset of ID8 tumours also had ID17, reportedly associated with a somatic TOP2A-K743N mutation[3]. ID1, ID2, ID5 and ID8 were found to be correlated with patient age at diagnosis, suggestive of a replication-based mechanism[3]. The causes of the remaining nine IDs are unknown. Several proposed channels were not informative across the cohort, and thus further optimization is required. Alternative methods of classifying indels have not been tested and may reveal biological insights not captured by this approach.

Rearrangement signatures. Another important class of somatic mutation is structural variation or rearrangements, which may delete, duplicate and/or reassemble relatively large chunks of chromosomal material in any orientation at kilobase to megabase scales. With use of a non-negative matrix factorization framework[18], 32 classification channels were proposed for putative RSs extracted from a local analysis of 560 WGS breast cancers[24]. The channels took into account how the rearrangement breakpoints were regionally clustered, the rearrangement type (for example, deletion, tandem duplication (TD), inversion or translocation) and the rearrangement size (FIG. 1b). Three of the six identified RSs correlated with tumour HRD: cancers with *BRCA1* but not *BRCA2* mutations displayed high numbers of RS3 small TDs (less than 10 kb), whereas cancers with *BRCA1* or *BRCA2* mutations showed a substantial number of RS5 deletions (less than 10 kb). The cause of RS1 long TDs (more than 100 kb), also associated with HRD, was not known[24]. The number of RSs was recently extended to 15 (REF.[4]). The 32-channel classification scheme[24] has also been used to report signatures in liver and ovarian cancer cohorts[25,26].

Recently, using a hierarchical Dirichlet process, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Structural Variation Working Group reported 16 RSs in a global analysis of ~2,559 WGS primary cancers, involving ~150,000 structural variations[5]. In addition to customary rearrangement classes, the study authors incorporated compound structural variation configurations including local *n*-jumps and chromoplexy, yielding 45 channels[5]. Two of the most prevalent structural variation classes, deletions and TDs, were further divided by size, replication timing domains and occurrences at fragile sites. The primary classification system was therefore highly complex.

Three signatures characterized by small, medium and large deletions emerged from the analysis[5]. A small deletion signature comprising mainly deletions of less than 10 kb and reciprocal inversions of less than 100 kb resembled RS5 deletions seen in *BRCA1*-mutated or *BRCA2*-mutated cancers[24]. A large deletion signature (10 kb to 3 Mb) was reminiscent of a complex form of breast RS2, while the mechanism of the medium-sized deletion signature was unknown[5]. The study authors hypothesized that template switching activity may explain their complex signatures, although this awaits external verification.

Five TD signatures were identified, differentiated by size and replication timing[5]. Both early-replicating and late-replicating small TD signatures (less than 55 kb)[5] were associated with *BRCA1* inactivation as previously reported[24,26]. The early small TD signature[5] displayed microhomology at breakpoint junctions, and also featured templated insertions believed to be the footprint of DNA polymerase-θ (POLQ)-mediated end-joining activity[27–29]. A late-replicating small TD signature was reported as enriched with FANC gene mutations[5]. Overall, small TD signatures are often associated with *BRCA1* loss in breast and ovarian cancers, although this association is not seen in liver, lung and cervical cancers[5,30]. Distinct mutagenic processes could plausibly converge onto similar TD phenotypes in different tissues.

We are still in the earliest stages of understanding how to classify indels and rearrangements. The current multifarious channels used in ID and RS extraction do not enable comparisons across studies. The elegance of the mutational signature framework does not lie in the mathematical algorithms — often simply matrix decomposition approaches. Rather, it is in how mutations are classified before factorization. An excessive number of uninformative channels reduces the power for signature detection (FIG. 1b). Conversely, channels that are too few may reduce the likelihood of discerning new biology. Channels that are too complex will reduce usability and possibly result in mixed signatures, rendering interpretation unnecessarily challenging. Additionally, there are orders of magnitude fewer indels and rearrangements than single-base substitutions; thus, other potential issues relating to power may yet reveal themselves.

Copy number signatures. Discrete mutational processes can drive the gains and losses of DNA (that is, copy number alterations in cancers). To date, few copy number signatures have been reported in local analyses of ovarian, prostate and soft tissue cancers using different methods[31–33]. The classification of copy number features before extraction used distribution-based features and was complex and cohort specific[31,32]. Copy number can be inferred from low-pass shallow sequencing or microarray data and may be a less expensive method for tumour classification and disease outcome prediction. However, copy number signatures have a limited resolution, as they report genomic changes at chromosomal and not at nucleotide scale, and thus will not have the accuracy afforded by substitution and indel phenotypes.

Considerations for interpretation

What does a signature reveal biologically? A mutational signature is the outcome of a mutagenic process comprising some form of DNA damage, subsequently acted

Poly(dA:dT)
Homopolymeric stretches of deoxyadenosine (dA) or deoxythymidine (dT) nucleotides on one strand of double-stranded DNA. They are overabundant in eukaryotic genomes and constitute a hotspot of mutagenesis.

Non-negative matrix factorization framework
An unsupervised machine learning framework in which a matrix is factorized into (usually) two matrices, with the property that all three matrices have no negative elements, thus allowing us to model a data matrix as linear combinations of a set of basis vectors (building blocks).

Hierarchical Dirichlet process
A non-parametric Bayesian approach to clustering grouped data.

Local *n*-jumps
A cluster of *n* structural variants in a single genomic region, usually phased to a single derivative chromosome, exhibiting some copy number gains and junctions with inverted and non-inverted orientation.

Chromoplexy
Large chained and weaved genomic rearrangements that involves multiple chromosomes.

Template switching
A recombination-based mechanism in which stalled polymerase finds an alternative template such that it can 'borrow' to get around a DNA lesion on the damaged parent strand and restart replication. This is most commonly the newly synthesized daughter strand on the sister chromatid or other sequences with homology to the single-stranded DNA region.

Matrix decomposition
Also called matrix factorization, works by decomposing the user-item interaction matrix into a product of two lower-dimensionality matrices, such as $M \approx S \times E$, where M is the catalogue matrix, with mutation types as rows and samples as columns, S is the signature matrix, with mutation types as rows and signatures as columns, and E is the exposure matrix, with signatures as rows and samples as columns.

Strauss's A rule
The preferential incorporation of adenine opposite a non-instructional DNA blocking lesion or across an abasic site.

upon by DNA repair and/or replicative machinery. This definition, however, faces biological complexities that also limit mathematical analyses.

First, a single type of primary DNA damage could be acted on by more than one DNA repair or replicative pathway, resulting in disparate outcomes (FIG. 2a). For example, APOBEC deamination of cytosine to uracil (C>U) may be the initial insult. Uracil may enter the replication process uncorrected and pairs with A during normal DNA replication, resulting in C>T mutations that are characteristic of SBS2 (REF.[34]). Alternatively, uracil may be processed by uracil-DNA glycosylase (UNG) as part of the base excision repair pathway, resulting in a so-called apurinic/apyrimidinic (AP) site that does not contain a DNA base[35,36]. Abasic sites may undergo Strauss's A rule[37] to produce SBS2 C>T mutations, or the predilection of DNA repair protein REV1 for insertion of C opposite uninformative AP sites would result in C>G transversions; this would present as SBS13, of which C>G transversions are a key feature[34] (FIG. 2a).

Second, any given repair protein may have multiple functions and may act on different types of DNA damage. When a repair protein is defunct, multiple compensatory pathways may be activated to deal with various forms of DNA damage. Thus, a defect in a single gene such as *BRCA1* could cause multiple signatures because of the multifaceted role of BRCA1 and the multitude of compensatory repair pathways that are called upon in its absence[24] (FIG. 2b). Arguably, each signature should be considered as unique because different types of initial DNA damage are required to generate substitution or rearrangement patterns. Thus, attempts to perform mutational signature analyses by combining different mutation classes may seem mathematically novel but may not be biologically correct. Furthermore, having signatures comprising mixed sources of DNA damage and repair would handicap efforts to understand individual signatures mechanistically. It may be advisable to regard signatures of different mutation classes as independent readouts and seek collinearity going forward. We thus contend that mutational signatures should be kept to individual classes.

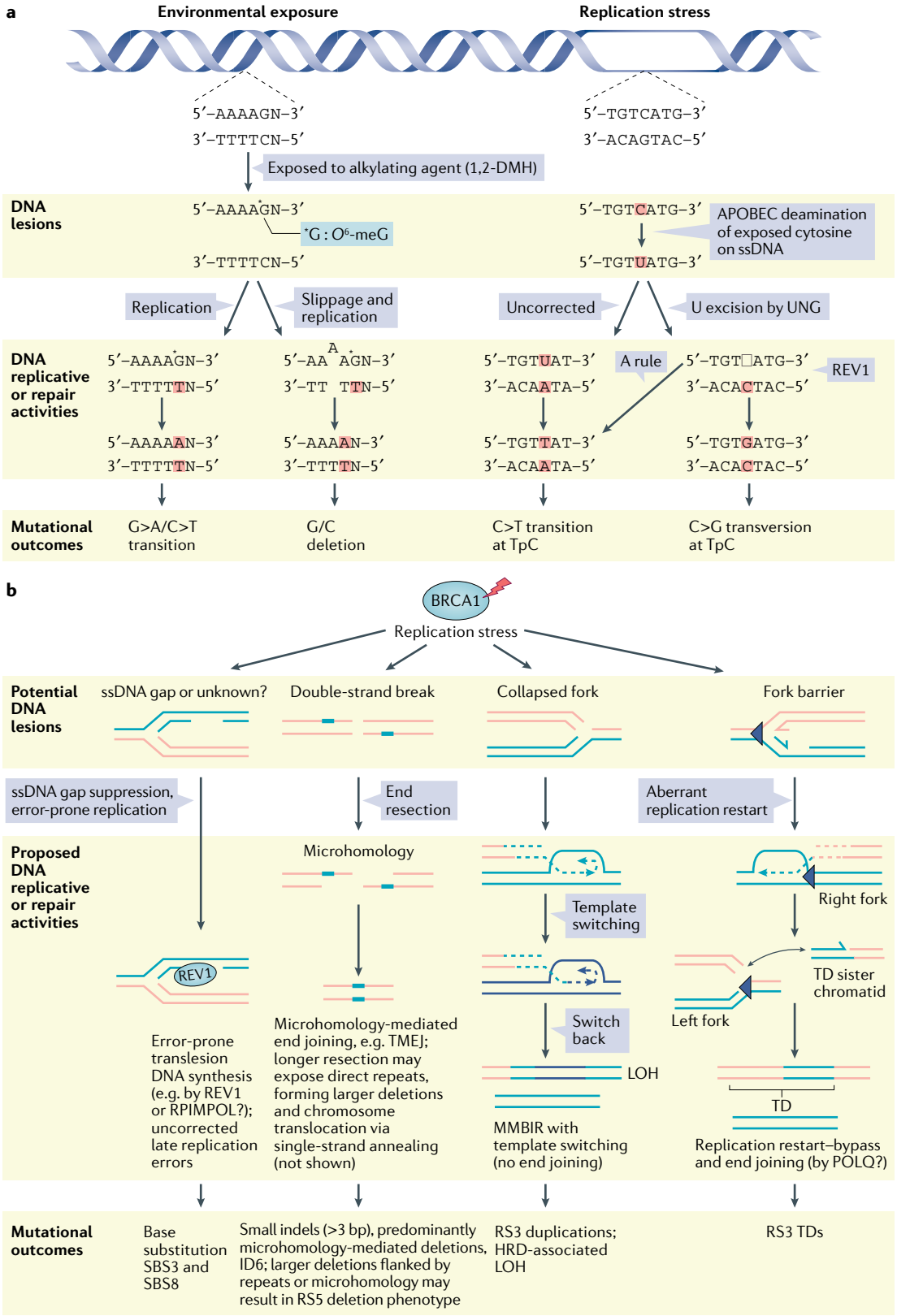
What signatures are present in my data set? Users of the mutational signature framework often want to know what signatures are present in their data set. To answer this question, they may seek to perform a fresh or de novo 'extraction' to identify the signatures before seeking the amount of each identified signature in each sample (a process known as 'assignment')[38]. An alternative is to rely on previous reported signatures or a predefined set of signatures, performing only the assignment step; this approach is known as 'signature fitting'[38].

The mutation profile of a sample is referred to as a 'mutational catalogue', *M*. It is a composite of all the mutagenic processes that were active at some point in the cancer cell lineage. It can be approximated as a linear combination of mutational signatures, *S*, each contributing a proportion of mutations (*E*) to the genome (that is, $M \approx S \times E$) (FIG. 3a). Thus, given a collection of mutational catalogues, matrix decomposition algorithms can infer the set of mutational signatures, *S*, that are recurrent across samples and estimate their relative contributory exposures, *E*, specifying how many mutations can be attributed to each signature in a sample[18]. Although intuitively simple, signature extraction using matrix decomposition techniques is a mathematical problem that may have multiple solutions (FIG. 3a). Deciding on the precise number of signatures present in a data set is not always straightforward and depends on the computational framework used. In an optimization framework[3,18], matrix decomposition is performed multiple times, and the resultant signatures are clustered. The number of signatures, and hence the robustness of the solution, is traded off against the error between the original mutational catalogue and the reconstructed mutational catalogue. In a statistical framework, the optimal number of signatures is learned on the basis of model selection or likelihood[39–42].

Drawing on published literature and our own expertise, we found that even if the number of signatures selected to represent a data set is one or two off the true number, common mutational signatures tend to be enduring in terms of their detection. Weaker signatures that are lower in frequency tend to suffer from being miscalled or mixed with other signatures. Mutational signatures with particularly prominent features such as tall peaks at specific trinucleotide sequence contexts are more likely to be extracted as a distinct signature, while signatures with flatter, non-distinctive profiles can be miscalled. Consequently, signatures do not have equal likelihoods of being extracted (FIG. 3b).

When signature extractions are being performed, multiple samples taken from different sites in the same patient or replicates of the same sample (biopsied multiple times) should probably not be presented for de novo extraction together, particularly in small cohorts, as this will introduce bias in detecting signatures among shared mutations. Another useful rule of thumb is to ask whether results from a new mutational signature extraction resemble those of previous extractions of a similar tissue type. If one is obtaining a large number of 'novel' signatures, then the possibility of power-limited analyses — due to a small number of samples or due to low numbers of mutations per sample (BOX 1) — is likely to be the explanation. The extraction step to identify *S* has often been performed recurrently by many groups in the field for many sample cohorts of many cancer types. Thus, a reasonable alternative is to bypass this step and simply ask which of the previously published signatures *S* are present in one's data set, as discussed next. Simply put, assume *S* is known and seek *E*, given *M*.

How much of each signature is present in my data set? To estimate the exposure (also termed 'contribution of a signature' or 'activity of a signature'), users often use one of several available software tools[4,14–19]. It is imperative to choose the most appropriate set of signatures to present to these tools. Users have variably used signatures that have been reported for a given tissue-type or have used a full list of reference signatures that have been reported across all tissues (for example, COSMIC mutational signatures v2 involving 30 signatures, recently updated to COSMIC mutational signatures v3.2 involving



53 signatures). At first glance, fitting all reported signatures across all tumour types may seem a good choice because it increases the likelihood of finding unusual signatures in one's data set (FIG. 3c). However, algorithms for estimating exposures are purely mathematical and will fit any and/or all presented signatures, including ones that are not biologically present. Critically, overfitting leads to misinterpretation of mutational signature data. This has led to the suggestion of a more conservative approach involving use of a tissue-specific signature set relevant to the new data set to overcome this 'overfitting' problem (FIG. 3c). Residual mutations that cannot

◀ Fig. 2 | **Mechanisms of signature generation. a** | Primary DNA damage from various sources can be resolved by disparate repair and/or replicative activities to produce multiple signature outcomes. Left: 1,2-Dimethylhydrazine (1,2-DMH) is an alkylating agent that causes primary damage to guanines, creating O^6-methylguanine (O^6-meG) particularly at ApG sites. The damaged G (now an O^6-meG) can pair with T, leading to a G>A or C>T substitution, or slippage can occur in addition to replication, resulting in a G or C insertion or deletion (indel). Right: Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) causes primary damage through deamination of deoxycytidines to deoxyuridines. Uracil-DNA glycosylase (UNG) removes uracil, and divergent reparative processes can occur from this point. Error-prone translesion polymerase REV1 has been postulated to insert cytosines opposite abasic sites to avoid detrimental replication fork stalling or collapse, resulting in C>G transversions at a TpC sequence context characteristic of single-base substitution mutational signature 13 (SBS13). Alternatively, uncorrected uracils and abasic sites that are not fixed via REV1 undergo contingency processing, for example via Strauss's 'A' rule. The outcome is C>T mutations at a TpC sequence context. **b** | BRCA1 is likely to play a role across diverse forms of DNA lesions that can lead to collapsed forks or double-strand breaks. When BRCA1 is not functional, a series of six signatures are observed as a result of these different types of damage that occur and the many compensatory repair pathways that are called upon to fix them. HRD, homologous recombination deficiency; ID, insertion or deletion signature; LOH, loss of heterozygosity; MMBIR: microhomology-mediated break-induced replication; POLQ, DNA polymerase-θ; RS, rearrangement signature; ssDNA, single-stranded DNA; TD, tandem duplication; TMEJ: DNA polymerase-θ-mediated end joining.

be fully explained by a given set of signatures can still be used for discovering new signatures[4]. Additionally, given the inherent variability between signatures, reporting the level of certainty associated with each signature's assignment in each sample can provide assurance regarding the accuracy of these assignments[4,19].

Another note of caution relates to the power available within a data set to perform mutational signature analyses and to robustly detect signatures (BOX 1). This is largely dependent on the type of sequencing experiment performed, as this determines the number of mutations that are detected, and on which such analyses are performed. Occasionally, hypermutator phenotypes such as those caused by APOBEC enzymes, MMR deficiency, DNA polymerase mutants or environmental agents could result in very high levels of mutagenesis and detectable patterns in an exome or a gene panel assay[20,43]. However, more often than not, the lack of power with these approaches relative to whole-genome sequencing will result in erroneous assignment of signatures to samples. This is important because once signatures and exposures have been obtained for a set of samples, the next step is often to draw biological and clinical conclusions. If the assignment is incorrect, so too will be the interpretation.

Association, not causation. To gain biological insights, we frequently associate mutational signatures with factors such as driver mutations, germ line variation, epigenetic modifications, and environmental or therapeutic exposures. However, these remain associations until causation is proven.

For example, SBS1 (REF.[2]), characterized by C>T transitions at methylated CpG dinucleotides, may be caused by spontaneous deamination of 5-methylcytosine, enzymatic deamination of cytosine or polymerase errors[44]. The burden of SBS1 is associated with the age at diagnosis in virtually all tumour tissues examined[2,7]. This translates neither to ageing being the cause of the signature nor to the signature being the cause of ageing. It is merely an early association, detected because the deamination of methyl-CpG dinucleotides happens spontaneously and continuously in all cells and is thus easily detectable[2,45,46]. When millions of cancers have been sequenced, there will probably be enough data points to show that many signatures show a correlation with age.

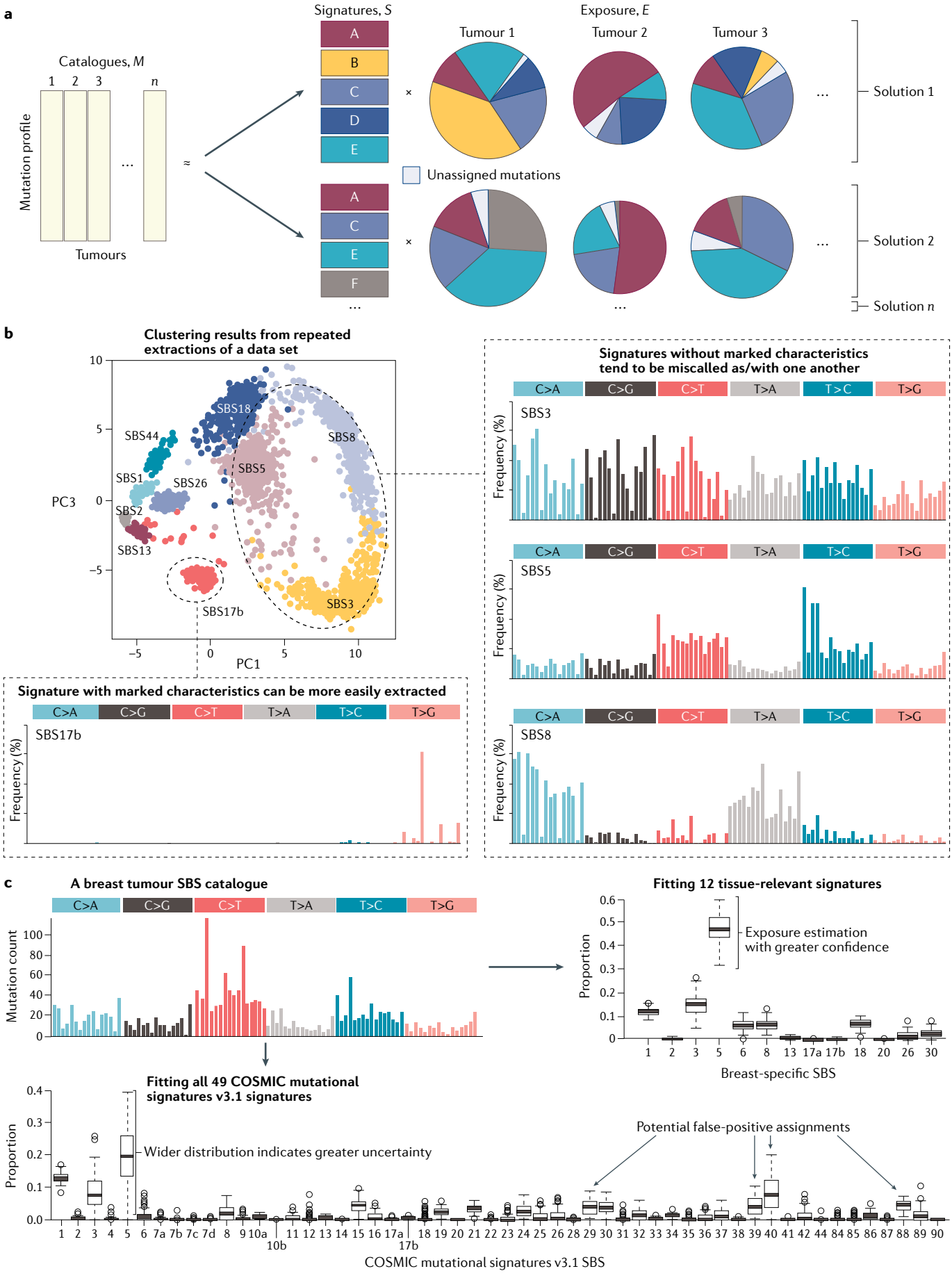
SBS1 is widely referred to as a clock-like signature[7], although it remains unclear whether this 'clock' refers to mutation accumulation in terms of cell division or time. For instance, in 1 year, a cell could divide ten times or 1,000 times. It is unclear whether the 'clock' refers to time regardless of the number of cell divisions or to the number of cell divisions regardless of time. Furthermore, the term 'clock' communicates a uniform rate, yet deamination is likely to vary over time or cell divisions. Analyses that use SBS1 to time cancer evolution sometimes assume that SBS1 occurs at a homogenous rate[7]. However, mutation acquisition per cell division may change as tissues evolve. In precancerous lesions, C>T substitutions at CpG dinucleotides can be approximately tenfold higher than in normal cells due to replication stress[47]. Thus, care must be taken when one is reporting an association as it could result in erroneous propagation of concepts and inappropriate use of signatures.

Another cautionary example is SBS17 (REF.[2]), characterized by T>G transversions at NTT and T>C transitions at CTT sequence contexts. This signature was first described in oesophageal and stomach cancers, and is present at a wide range of mutation densities ranging from low levels to hypermutator phenotypes; gastric acid exposure was raised as a potential cause[8,48]. Recently, it was noted in a wide range of metastatic cancers[49–51] and human small intestinal organoid cultures treated with 5-fluorouracil[52]. However, the notion that SBS17 may be a direct consequence of gastric acid or 5-fluorouracil exposure becomes less likely when we consider that the signature has also been observed in unexposed tissues such as breast cancer[24], and in untreated immortalized mouse embryonic fibroblasts[53,54]. Indeed, treatment with bile acids and gastric acid, two components of refluxate and risk factors for oesophageal cancer, was shown to increase the levels of 8-oxo-dGTP in oesophageal tissues and cell lines[55]. In vitro experiments have also demonstrated that oxidized guanine in the nucleotide pool can cause T>G transversions when (mis)incorporated opposite A on the template strand with subsequent insertion of dCTP during the second round of replication[56]. Collectively, these data suggest that T>G mutations of SBS17 are possibly by-products of oxidative damage in the nucleotide pool, which may be secondary to exposure to gastric acid or 5-fluorouracil. Why the misincorporation occurs at specific trinucleotides, however, remains unclear.

Like SBS17, SBS18 is likely to be an endogenous signature that can be secondarily amplified under cellular stress regardless of the origins of the primary stressors (FIG. 4). Attributed to DNA damage from oxidative species, SBS18 is most likely due to 8-oxo-dG[57–59]. Therefore, in vivo or in vitro, any changes or physiological

Replication stress
The slowing or stalling of replication fork progression and/or DNA synthesis in response to DNA damage or any hindrance to DNA replication.

8-Oxo-dGTP
8-Oxo-2′-deoxyguanosine 5′-triphosphate, the oxidized form of 2′-deoxyguanosine 5′-triphosphate (dGTP). It can mispair with A, leading to C>A/G>T transversion mutations.


a
Catalogues, M
1 2 3 n
Mutation profile
Tumours
≈
Signatures, S
A
B
C
D
E
A
C
E
F
Unassigned mutations
Exposure, E
Tumour 1
Tumour 2
Tumour 3
Solution 1
Solution 2
Solution n
b
Clustering results from repeated extractions of a data set
SBS18
SBS44
SBS8
SBS1
SBS26
SBS5
SBS2
SBS13
SBS17b
SBS3
PC3
PC1
Signatures without marked characteristics tend to be miscalled as/with one another
C>A C>G C>T T>A T>C T>G
SBS3
SBS5
SBS8
Frequency (%)
Signature with marked characteristics can be more easily extracted
SBS17b
c
A breast tumour SBS catalogue
Mutation count
Fitting 12 tissue-relevant signatures
Proportion
Exposure estimation with greater confidence
Breast-specific SBS
Fitting all 49 COSMIC mutational signatures v3.1 signatures
Wider distribution indicates greater uncertainty
Potential false-positive assignments
COSMIC mutational signatures v3.1 SBS

◀ Fig. 3 | **Challenges associated with mutational signature frameworks. a** | Performing a de novo extraction of mutational signatures from a new data set can result in multiple potential solutions. Starting with a catalogue of mutations, *M*, from multiple tumours, the purpose of the exercise is to identify the set of signatures, *S*, and the amount of each signature or exposures, *E*, per tumour in the data set. Two possible solutions are depicted. **b** | Signatures do not have equal likelihood of being extracted. The figure shows a principal component (PC) analysis plot of mutational signatures extracted 400 times, each with ten resultant signatures ($k = 10$, the near optimal number), from 560 cancer whole genomes[24]. Well-defined clusters signify robustly extracted signatures, whereas less discretized clusters that interconnect indicate uncertainty in the signatures extracted across repeated analyses. Clusters were identified by the signature most similar to the cluster medoid. **c** | Signature fitting of a breast tumour substitution catalogue (PD7319a)[4]. Signature assignment depends entirely on given *S*. Fitting too many signatures (bottom, COSMIC mutational signatures v3.1) leads to increased levels of uncertainty in exposure estimation and potentially false-positive assignments. Restricting *S* to a selected set of signatures on the basis of previous knowledge of the samples (that is, breast-specific signatures) reduces the likelihood of overfitting and provides more accurate exposure estimation (top right). Here, bootstrapping was performed to produce a distribution of exposures (boxplots) across 100 fits, providing a confidence measure for signature exposure values fitted in a sample[4]. Boxes show median, first and third quartile, with whiskers extending at most 1.5 interquartile range. SBS, single-base substitution mutational signature. Part **b** adapted with permission from REF.[3], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

manipulations that increase the amounts of oxidative species, including 8-oxo-dG, will increase the amount of SBS18 in human cells. This probably explains why it is seen in some tissues in vivo but is also a prominent feature of cultured cells[20,60,61] and can be induced by various environmental agents[13].

Thus, we need to be cognizant of potential confounders and how the overall message is communicated with regard to associations within genomic data. A mutational signature may be the primary, direct outcome of a particular process; it could also be an indirect, secondary consequence of multiple potential processes (FIG. 4). As exemplified by SBS17 and SBS18, even endogenous signatures can be brought on or amplified by exogenous exposures.

Timing of mutational signatures

Precancer. Recent sequencing forays into normal and precancerous skin, oesophageal, colon, endometrial, lung and bladder tissues have revealed mutational signatures caused by deamination of 5-methylcytosine, APOBEC enzymes and UV exposure[23,62–67]. These findings reinforce how mutational processes may precede tumorigenesis and may be tolerated in human cells (BOX 2; FIG. 5). Whole-genome sequencing assessments of pluripotent stem cells have also revealed extensive UV-related mutagenesis in genomes of fibroblast-derived cellular models[68]. Indeed, nearly three-quarters of stem cell lines studied had mutation burdens that rivalled cancers[69]. Thus, the extent of mutagenesis in normal and precancerous tissues challenges the prevailing view that mutagenesis is pathognomonic of cancer since heavy mutagenesis can be found in normal, healthy cells[70] (BOX 2). The challenge for the research community is to discern mutational signatures that have clinical relevance (that is, those that may indicate biological abnormalities that are potentially targetable or that may serve as predictors or prognosticators). This can become clearer only through sequencing endeavours that have accompanying treatment and outcome data if we are to identify genomic features that differentiate indolent tumours from aggressive tumours.

Metastatic cancer. Most of the mutational signatures reported to date have been derived from primary cancers. Primary tumours report the earliest mutational processes during tumorigenesis, which may be relatively benign and not life-limiting (FIG. 5). Mutational signatures associated with metastatic cancers have been relatively understudied, with most previous efforts focused on small tumour-specific cohorts using gene panel approaches or whole-exome sequencing[49,50,71–75]. In a large-scale whole-genome sequencing metastatic breast cancer study, nearly all mutational signatures reported in primary cancers were observed in metastases[49]. However, the relative contribution of mutational signatures differed, with APOBEC-driven processes enriched in metastases relative to primary breast tumours. Other metastatic cancer studies have likewise reported relative enrichment of APOBEC signatures and SBS17 of uncertain cause[50,76–79]. SBS9, which was rarely found in primary colorectal cancers, has been observed in metastatic colorectal cancers[79]. A recent pan-cancer study on ~2,500 WGS metastases across multiple cancer types suggested that metastatic cancers were enriched in endogenous signatures associated with repair deficiencies[51]. Patients presenting with metastases some time after their primary diagnosis may have been exposed to various therapeutic regimens, and this can be reflected in their mutational signatures[78,79]. As an example, signatures caused by platinum exposure have been reported in various cohorts with advanced and metastatic cancers[9,51,75,78,79].

Studies in metastatic cancers could provide useful insights into therapeutic resistance mechanisms and could reveal signatures that forecast a poor prognosis. However, to fully interpret metastatic cancer mutational data, sequences of matched primary cancers are required to distinguish mutations of the founding clone from those that have driven dissemination (FIG. 5). To clinch the next level of understanding, it is imperative to obtain all relevant samples and collate clinical data, including timelines of exposures and outcomes. Signature analyses mapped to tumour evolutionary paths could help us to understand trajectories of mutational processes more effectively, informing modifications to treatment strategies[80].

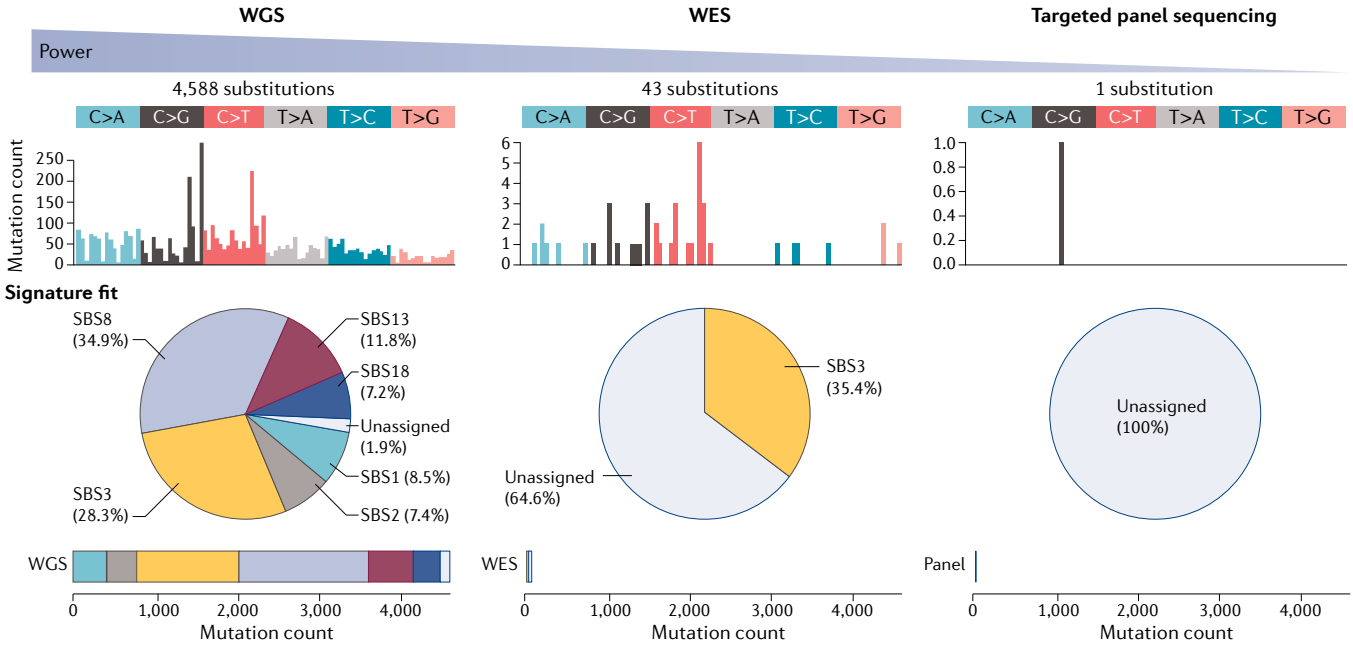
Experimental validation

Thus far, mutational signatures have been derived analytically through mathematical decomposition of complex patient-derived cancer mutation data. In recent years, there have been efforts to validate them experimentally.

Mutational signatures of environmental agents. A comprehensive, systematic in vitro screen of 79 International Agency for Research on Cancer (IARC) class I and class IIa/IIb known or suspected environmental carcinogens was presented recently in a human induced pluripotent stem cell system[13]. A total of 53 SBSs were observed from 41 agents tested. Well-studied environmental mutagens such as UV light, tobacco smoke and aristolochic

Box 1 | **Power for signature detection with different sequencing approaches**

In general, whole cancer genomes have thousands of mutations, hundreds of insertions or deletions (indels), and tens to hundreds of rearrangements. By contrast, exome-sequenced cancers have only ~1–3% of the human genome footprint (depending on the bait set used), and therefore orders of magnitude fewer mutations. In this example of a breast cancer genome (PD6413a)[4], whole-genome sequencing (WGS) at 40-fold depth revealed 4,588 substitutions. A whole-exome sequencing (WES) experiment with the same sample revealed 43 mutations. Given that a substitution signature profile has 96 elements, when 43 mutations are distributed across 96 channels, the numbers of mutations may be so low as to result in many channels with counts of 0 or 1. The power to reliably discern signatures then becomes limited (pie charts). Targeted sequencing has an even smaller genomic footprint, and here a typical TruSight Oncology 500 (Illumina) targeted panel would have detected one substitution, further reducing the reliability of signature assignments. For WES and targeted sequencing data, indels may be down to single digits, and rearrangements would not be reported at all.



SBS, single-base substitution mutational signature.

acid exhibited greater effect sizes than many other mutagens[13]. This is noteworthy because although this compendium of mutational signatures provides a reference set to hunt down exogenous exposures in human cancers going forward, signatures with small effect sizes may be difficult to detect in human tumours where multiple endogenous and exogenous mutational processes are operative.

This screening exercise also produced ten IDs and eight DBSs[13]. Deeper exploration revealed novel mechanistic insights. Analyses of surrounding sequence context of indels generated by cisplatin and polycyclic aromatic hydrocarbons suggested that error-prone base excision repair is involved in reparation near, but not at, primary adducts created by these mutagens[13]. Additionally, repair of double substitutions often follows Strauss's A rule[37], producing an NN>TT outcome, irrespective of mutagen type. This comprehensive screen demonstrated how single primary adducts could cause multiple signatures because of disparate repair pathway activity (FIG. 2a). Likewise, diverse mutagens could generate similar signatures; for example, dibenzo[*a*,*l*]pyrene diol epoxide (a polycyclic aromatic hydrocarbon in tobacco smoke) and aristolochic acid are unrelated compounds, yet both produced strikingly similar T>A or A>T components in their signatures[13].

Genotoxin screens in model systems. Non-human models that are more permissive for mutagenesis, including *Caenorhabditis elegans*, chicken DT40 cells and mouse embryonic fibroblasts, have been used to recapitulate mutational signatures[81–87]. Model organisms such as *C. elegans* and *Saccharomyces* may offer compact genomes, convenient life cycles and relative ease of genetic manipulation and bottlenecking. However, they can produce signatures different from those seen in humans[13,85,88].

Organoids have been used to demonstrate novel causes of signatures. For example, human intestinal organoids subjected to repeated luminal microinjections of colibactin-producing *Escherichia coli* showed a unique substitution signature (SBS88) characterized by T>A and T>C mutations particularly at ATA, ATT and TTT motifs, and an ID (ID18) featuring single A or T deletions at poly(dA:dT) tracts[89]. These findings were substantiated by molecular dynamics simulations and experimental data from an independent study showing enrichment of colibactin-induced damage at (A+T)-rich hexameric sequence motifs[90]. Similar colibactin-damage patterns were also seen in colorectal cancers and healthy colorectal epithelial cells[23]. Together, these data provide strong evidence of mutational signatures caused by a secondary metabolite secreted by certain strains of bacteria[89–91].

Collectively, these pioneering studies underscore the vulnerability of human DNA to exogenous insults and provide reference frameworks for environmental genotoxins.

Mutational signatures of gene edits. Experimental exploration of endogenous signatures has been relatively restricted. More technically challenging to accomplish, they can be explored through knock-ins, knockdowns or knockouts of various genes, or through overexpression of oncogenic proteins or processes that drive tumorigenesis.

To establish experimental and analytical proof-of-principle methods, a near-haploid cancer cell line, HAP1, was used for a CRISPR–Cas9-mediated knockout screen of nine DNA repair or replication genes[92]. Knockout of *EXO1*, *FANCC* or *MSH6* in HAP1 cells produced multiple mutational patterns of substitutions, indels and/or rearrangements. The other gene knockouts did not produce detectable signatures under the experimental conditions used in the study[92].

Other model systems have been used for gene editing, primarily involving loss-of-function mutations in tumour suppressor or DNA replicative/repair pathway genes, to study mutational signatures. In organoids derived from adult mouse liver stem cells hemizygous for *Ercc1*, increased base substitutions were ascribed to SBS8 (REF.[93]). The same study also claimed increased SBS8 mutations in *XPC*-knockout human intestinal organoid cultures[93]. Knockout of *NTHL1* in human colon organoids produced a signature identical to SBS30 (REF.[94]), first described in breast cancer[24]. Subsequent examination of the germ line revealed a heterozygous nonsense mutation in *NTHL1*, causing a premature stop codon with loss of heterozygosity of the wild-type allele in the breast tumour[24,94].

Latterly, a systematic study of the mutational characteristics of in vitro knockouts of 42 DNA repair or replication genes was performed in human induced pluripotent stem cells[59]. Critically, cells were unchallenged by exogenous mutagen exposure, permitting identification of genes that are fundamentally important in guarding the genome against natural sources of DNA damage. Reassuringly, different mutant genotypes of genes including *OGG1*, *UNG*, *EXO1*, *RNF168*, *MLH1*, *MSH2*, *MSH6*, *PMS2* and *PMS1* each produced marked mutational signatures. Most knockouts showed modest effects (twofold to fivefold increase above the background mutation burden), except for MMR genes, where knockout resulted in up to a tenfold increase in mutation burden. This study emphasized the value of detailed signature analyses and revealed insights into mutational mechanisms, such as how OGG1 and MMR proteins coordinate to sanitize oxidized guanines at specific trinucleotide motifs[59]. The study authors also postulated that the modest T>A peaks at ATT or TTA motifs in MMR deficiency signatures are due to "reverse template slippage" events at adjoining A and T homopolymers.

While cellular model systems may reveal some causes of new signatures, it is envisaged that they may also help to clarify areas of uncertainty. For example, seven substitution signatures (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44) were reportedly associated with MMR deficiency, but it is not clear how these signatures reflect specific MMR defects[3]. An independent analysis of 3,107 WGS primary cancers from various organs identified two MMR deficiency-associated signatures (RefSig MMR1 and RefSig MMR2), although variations

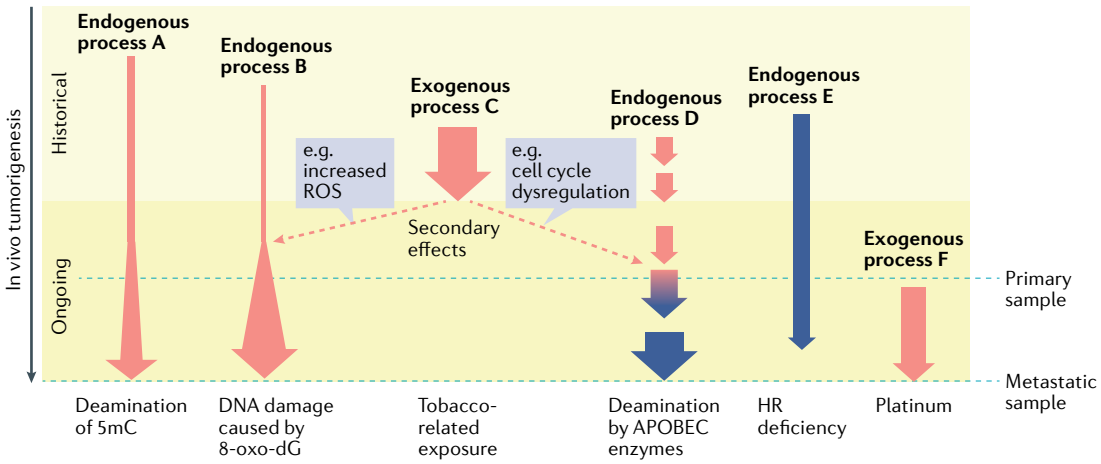
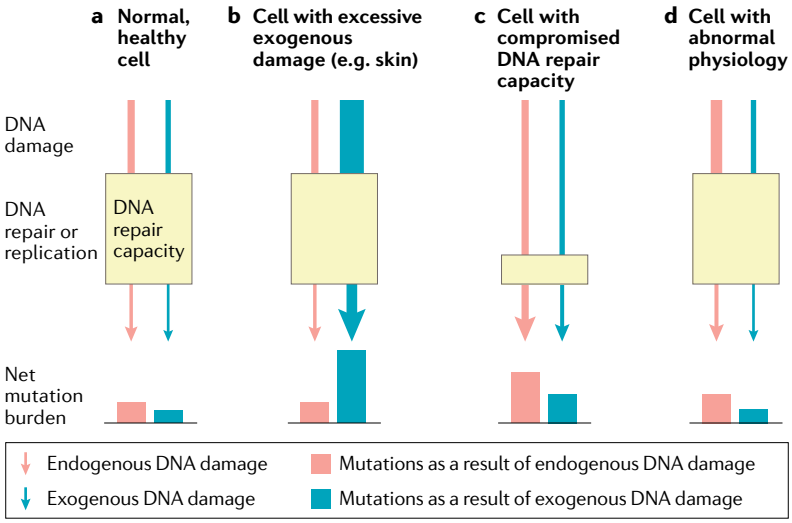


Fig. 4 | **An update on the concept of mutational signatures.** Mutational signatures are the imprints of various endogenous and exogenous mutational processes (labelled 'A' to 'F'). Some processes are historical, while others are ongoing and even intermittent (process D). Mutational processes that cause signatures in a direct manner can be considered primary signatures. There may also be augmentation of certain signatures secondary to cellular abnormalities that arise due to primary exogenous mutagen exposure (red dashed arrows). Some mutational processes may be clinically informative (highlighted in dark blue); for example, process D, which when amplified may signal dysregulation of the cell cycle, or process E, which may indicate a deficiency of a DNA repair pathway that has synthetically lethal interactions with particular therapeutic agents. Process F is an example of a late-onset iatrogenic exposure due to treatment. The horizontal turquoise dashed lines indicate different sampling times. APOBEC, apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like; HR, homologous recombination; 5mC, 5-methylcytosine; 8-oxo-dG, 8-oxo-2′-deoxyguanosine; ROS, reactive oxygen species.

Box 2 | **Mutagenesis is normal**

Contrary to the notion that cancer genomes are laden with mutations, while normal genomes are unaltered, it is now clear that normal cells can carry a substantial amount of mutagenesis, including marked mutational signatures of endogenous and exogenous sources[23,62–67]. This is because human cells experience a constant stream of DNA damage from endogenous (see the figure, pink arrow) and exogenous (see the figure, blue arrow) sources. Although our cells have a complex set of replicative and repair pathways available (see the figure, yellow box) to sanitize the genome, unrepaired damage may persist (see the figure, arrowhead), resulting in a net mutation burden (see the figure, pink and blue bars), even in normal, healthy cells (see the figure, panel **a**). Some cells are subjected to a large amount of DNA damage, for example skin cells to UV light, and can show marked mutagenesis as a result of this exogenous exposure (see the figure, panel **b**) even though DNA repair capacity is normal[62]. It has been postulated that there are limits to the amount of DNA damage that our cells are permitted to repair[70]. This restraint is believed to prevent the DNA repair system from consuming too much of the cellular resource to maintain the genome. If the vast majority of DNA damage is non-deleterious because the genome is largely intronic and intergenic, it may not be necessary to fix all DNA damage. If a cell exposed to a large amount of external damage was engaged in repairing all DNA damage, it could risk cell death. In essence, the cell sacrifices genomic perfection and prioritizes cellular survival, resulting in permissiveness regarding mutagenesis in normal cells[70]. Under circumstances where DNA repair capacity is compromised, however, even normal amounts of DNA damage can amount to marked mutagenesis (see the figure, panel **c**). While many genes are involved in DNA repair, only a handful of them are truly critical for safeguarding the genome; thus, their impairment can cause direct, marked mutagenesis[59]. Mutagenesis can also arise when there is a change to cellular physiology despite no change to DNA repair capacity (see the figure, panel **d**); for example, a state of increased replication stress or increased production of reactive oxygen species. In these circumstances, mutational signatures that are endogenous are augmented as a secondary effect.



in the signatures were seen in different tissues[4]. In vitro, knockout of *MSH2*, *MSH6* or *MLH1* in human induced pluripotent stem cells produced substitution signatures nearly identical to RefSig MMR1, dominated by an excess of C>T mutations with a pronounced C>A peak at CCT, and IDs with prevailing T deletions at increasing lengths of poly(T) tracts as well as minor contributions of T insertions and C deletions[59]. By contrast, knockout of *PMS2* produced a signature dominated by T>C mutations with a small contribution of C>T mutations and a dwarfed C>A peak at CCT similar to RefSig MMR2 (REF.[4]). An ID with disparate proportions of T insertions and deletions was also noted as differing between knockout of *MSH2*, *MSH6* or *MLH1* and knockout of *PMS2*. Together, these experimental data validated gene-specific characteristics underlying human cancer-derived MMR deficiency-associated signatures, thereby revealing two mutational signatures consequent upon the abrogation of key MMR proteins in human stem cells[59].

Future experimental directions. Experimental data are thus beginning to emerge. Early analyses highlight the importance of experimental approaches to facilitate understanding of mutational mechanisms[12]. Further work is envisaged, including combinations of DNA damage sources on various DNA repair-deficient backgrounds, potentiation through non-genetic methods (for example, small-molecule inhibitors or enhancers) and exploration of transient bursts of mutagenesis. Critically, interpretation of experimental data is entirely dependent on sensible application of analytical principles. In vitro and in vivo data should be assessed separately from clinical data; they should not be aggregated in primary analyses but should be compared with clinical data in an independent manner. This constitutes true, independent validation of signatures derived from human cancers and will help to prevent misinterpretation.

Signatures created through controlled methods could be used as reference signatures for public health purposes, assessing environmental exposure in high-risk areas by designing assays to screen at-risk populations. However, it is important to note that the detection of mutational signatures through such a surveillance programme would not necessarily equate to cancer detection — it simply reports exposure. Thus, thoughtful studies are required before consideration of environmental signatures as potential early detection cancer markers.

Clinical applications

To take mutational signatures towards clinical applications, it is prudent to distinguish signatures that are clinically relevant from those that are not, as some studies report early mutational processes of normal cells that may not be life-limiting[95] (FIG. 4).

Homologous recombination deficiency. In 2005, synthetic lethality between poly(ADP-ribose) polymerase (PARP) inhibition and HRD was described, which propelled use of PARP inhibitors as a strategy for targeting *BRCA1*-deficient or *BRCA2*-deficient tumours[96–98]. Because chromosomal abnormalities and large-scale genomic losses were noted features of these tumours, the principle of 'genomic scars' was developed as a potential biomarker[99]. Assays such as the myChoice HRD assay (from Myriad Genetics), which measures loss of heterozygosity[100], telomeric allelic imbalance[101] and large-scale state transitions[102], have been developed as a companion diagnostic for PARP inhibitors to identify tumours with HRD[103].

Recently, it was demonstrated that *BRCA1* or *BRCA2* nullness gives rise to six genomic signatures, including two SBSs, one ID, two RSs and genome-wide loss of heterozygosity[24] (FIG. 2b). These multiple signatures are the direct consequences of homologous recombination pathway abrogation, and when combined provide exquisite sensitivity and specificity as a composite assay[104]. A mutational signature-based algorithm called

Synthetic lethality
Interaction between two genes when the perturbation of either gene alone is viable but the perturbation of either gene results in cell death.

'HRDetect' was derived, which showed remarkable sensitivity of 98.7% (area under curve 0.98) in predicting biological *BRCA1* or *BRCA2* deficiency in a substantial proportion of patients with breast or ovarian cancer (~22% and ~63%, respectively)[104]. Latterly, HRDetect was applied in an independent retrospective study and demonstrated association with platinum response in advanced breast cancer[105]. Furthermore, in high-risk patients with familial breast cancer, HRDetect offered a more robust assessment of HRD than germ line status and other genomic scar predictors, impacting selection for platinum-based or PARP inhibitor therapy[106]. Built on the success of HRDetect, alternative tools have also been developed for HRD detection; CHORD used the relative counts of different mutation types and features instead of mutational signatures as predictive features to detect HRD[107], while SigMA was designed to detect HRD-associated SBS3 in samples with low mutation counts or from targeted gene panels[108]. These newer tools await further validation.

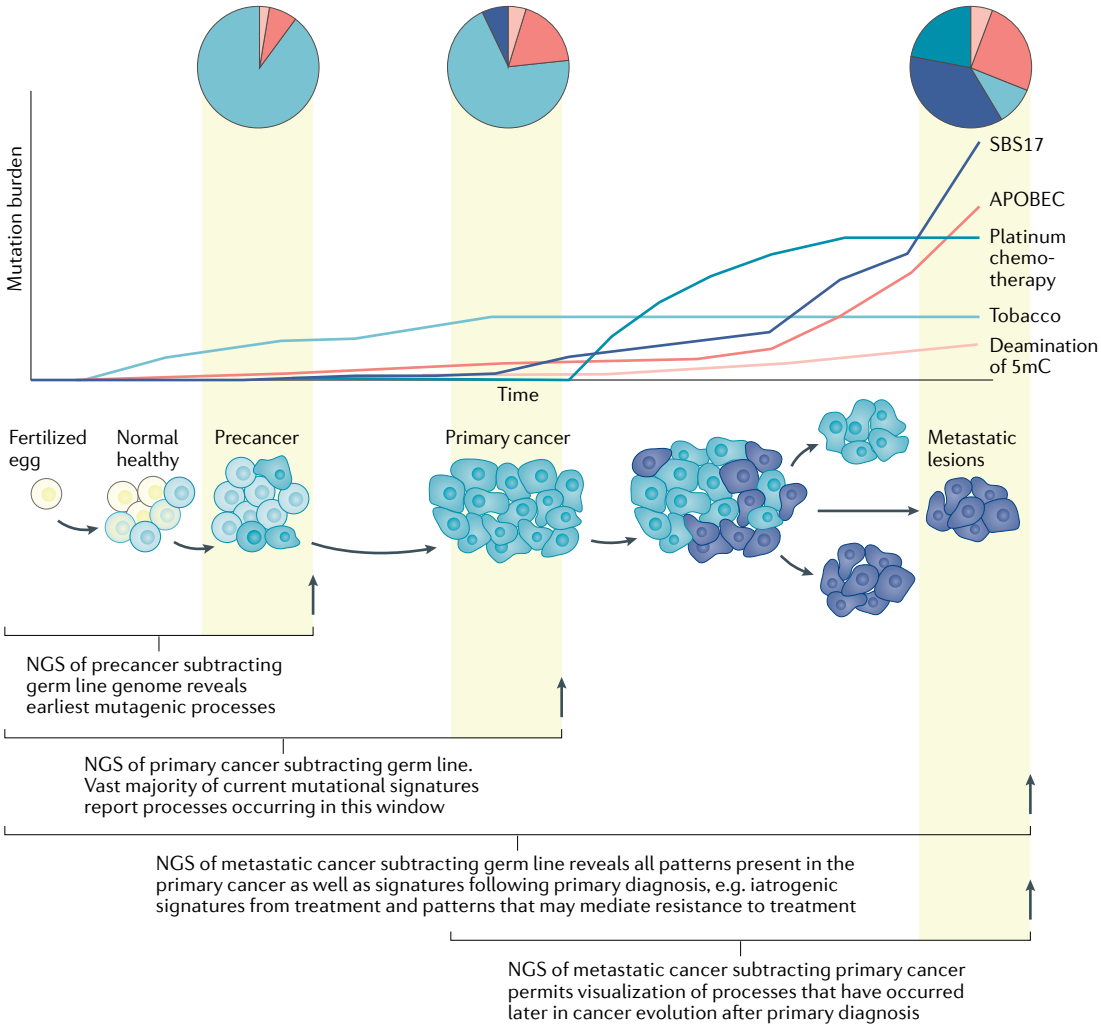


Fig. 5 | **Dynamics of mutational signatures over cancer evolutionary time.** Schematic representation of how a polyclonal population of cells can evolve into a dominant precancer clone, and then a frank carcinoma before metastasis. The graph shows how the mutation burdens of five different mutational signatures change over time. Single-base substitution mutational signature 1 (SBS1) (pink line; caused by deamination of 5-methylcytosine (5mC)) is accumulating in a linear way in the early stages. Tobacco exposure (pale blue) from early adulthood leaves a marked mutational pattern with high mutation counts. However, there is cessation of accumulation of this mutational signature when the patient stops smoking in mid-adulthood. Before formation of the precancerous neoplasm, there is an increase in apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC)-related mutagenesis (red). Sequencing of the precancerous lesion and subtraction of the germ line would reveal the earliest mutational processes in tumorigenesis (leftmost pie chart). At the point of primary cancer diagnosis, there is a modest increase in APOBEC signatures and the presence of SBS17 (blue line, middle pie chart). When the patient presents with metastatic disease, the sequencing of this disseminated lesion and subtraction of the germ line will reveal all the processes that have contributed to the metastases, including all the processes present in the primary cancer. By contrast, subtraction of the primary cancer could reveal insights that are specific to the post-primary window, including iatrogenic signatures from treatment (for example, dark turquoise line; caused by exposure to platinum chemotherapy), immune-genomic patterns and signatures that mediate treatment resistance. NGS, next-generation sequencing.

In a real-world clinical diagnostic population-based setting, HRDetect was applied to 254 cases of triple-negative breast cancer (TNBC)[109]. Patients with a high HRDetect score receiving adjuvant chemotherapy showed improved invasive disease-free survival (hazard ratio 0.42, 95% confidence interval 0.2–0.87) and distant relapse-free interval (hazard ratio 0.31, 95% confidence interval 0.13–0.76) compared with patients with a low HRDetect score[109]. Critically, this was irrespective of whether a genetic or an epigenetic cause for HRD was identifiable; indeed, drivers of HRD were detected in only approximately two-thirds of the cohort with a high HRDetect score, attributable to germ line or somatic *BRCA1* or *BRCA2* mutations, promoter hypermethylation of *BRCA1* or *RAD51C*, or biallelic *PALB2* loss. In other words, the mutational signature approach was able to prognosticate outcomes in cases that would be missed by conventional targeted sequencing assays, up to ~30% of tumours with HRD signatures[109].

Encouragingly, in the proof-of-principle phase II window clinical trial named 'RIO' (EudraCT 2014-003319-12), HRDetect robustly identified sporadic TNBCs with PARP inhibitor sensitivity[110]. Forty-three patients with treatment-naive TNBC were treated with the PARP inhibitor rucaparib for 2 weeks before surgery or standard chemotherapy. Cases with a high HRDetect score were correlated with HRD by functional RAD51 focus formation assays, and showed evidence of sensitivity to the PARP inhibitor as reflected in reduced circulating tumour DNA levels in patients with TNBC[110]. Compared with copy number-based HRD scars, HRDetect proved more specific in identifying TNBCs with functional homologous recombination defects in this small cohort of patients with treatment-naive tumours, suggesting that mutational signature assessment may more accurately identify cancers that would potentially benefit from treatment with a PARP inhibitor as a first-line agent.

Mismatch repair deficiency. MMR deficiency is associated with elevated mutation rates at short tandem repeats or microsatellite instability (MSI)[111–113]. Mutations in the MMR genes *MLH1*, *PMS2*, *MSH2* and *MSH6* and promoter hypermethylation of *MLH1* are predominant causes of MSI[114]. Sensitivity of MMR-deficient tumours to checkpoint blockade therapy[115–117] has spurred the development of multiple targeted or exome-based next-generation sequencing (NGS) assays to report MSI status[118–120]. The functional basis of sensitivity to checkpoint blockade is postulated to be due to an MMR deficiency hypermutator phenotype, which results in persistent creation of neoantigens, triggering long-lasting immunosurveillance, which is further enhanced by immune modulators such as anti-programmed cell death 1 (anti-PD1) and anti-cytotoxic T lymphocyte antigen 4 (anti-CTLA4)[121]. In 2017, the US Food and Drug Administration approved the checkpoint inhibitor pembrolizumab for any resectable or metastatic solid tumour with MMR deficiency, a first tumour-agnostic approval[122].

MMR-deficient tumours are thus worthy of identification irrespective of tissue origin. Current clinical assays for detection of these tumours range from immunohistochemical staining for concomitant loss of MMR protein pairs[123,124] to PCR-based assays to determine MSI (for example, a pentaplex assay to detect two mononucleotide markers – BAT25 and BAT26 — and three dinucleotide markers – D2S123, D5S346 and D17S250 (REF.[125])), and algorithms designed for NGS data, such as mSINGS[120], MSIsensor[119] and MSIseq[126]. Conventional electrophoresis methods suffer from insufficient sensitivity to resolve short indels (less than 3 bp), whereas most NGS-based assays primarily use targeted or exome-sequencing assays to determine MSI status on the basis of tumour mutational burden and/or MSI[119,120,126–128]. However, multiple biological abnormalities and environmental exposures can cause an increased burden of mutagenesis; thus, the crude tumour mutational burden measure may not be specific to MMR-deficient tumours. Furthermore, tumours that occur in tissues that do not have the high proliferative rates of colon or endometrial tissues can also demonstrate MMR deficiency, including in subclones[22], but may not have an equivalently high mutagenesis burden, making tumour mutational burden inadequately sensitive for MMR deficiency detection. A recent exercise leveraging experimentally generated signatures of MMR deficiency led to the development of an MMR deficiency classifier, termed 'MMRDetect'[59]. It uses substitution and indel mutational signatures associated with MMR gene defects to classify tumours, agnostic of mutation burden, and appears less likely to generate false-positive and false-negative calls. Validations of these assays are required in basket clinical trials to gauge their true relative merits.

Signatures associated with polymerase dysregulation. Another mutational signature associated with sensitivity to checkpoint inhibition is SBS10 (REF.[2]), which is caused by POLE proofreading defects[129,130]. The mechanism underpinning this sensitivity is unclear but is postulated to be neoantigen formation. SBS10 is strikingly distinctive and often results in heavy mutagenesis. While there have been anecdotal reports of responses to checkpoint blockade in tumours with SBS10 in atypical tissue types with which one does not typically associate *POLE* as a potential driver[131,132], systematic studies are required to understand pan-cancer benefits of using this signature as a predictive indicator of therapeutic sensitivity.

Recently, IDs with 1-bp insertion bias were reported in cancers with defective proofreading polymerases (MS-sig2 and MS-sig4)[133]. The tumours were termed replication-repair-deficient, and a POLEness score based on normalized MS-sig2 insertions was used to distinguish responders from non-responders in a small cohort of patients receiving immune checkpoint inhibition therapy. Other signatures that are present and hypothesized to be due to isolated polymerase dysregulation (for example, SBS28) or concomitant with MMR deficiency (for example, SBS14 and SBS20)[3] may turn out to be predictors of therapeutic response in due course.

APOBEC-related mutagenesis. APOBEC-related SBS2 and SBS13 are ubiquitous and well described in multiple tumour types[1,2,134]. Mutational densities differ widely,

Microsatellite instability (MSI). Variability in the length of base pair repeated sequences (less than 5 bp) due to short insertions/deletions caused by replication slippage and that is normally kept stable by mismatch repair.

with some samples showing low levels of APOBEC signatures, whereas others, such as cervical and bladder cancers, demonstrate consistently elevated background levels[134]. Additionally, individual cases within many tumour types can show marked hypermutator phenotypes. Notably, APOBEC-related deamination requires single-stranded DNA as a substrate. As a cancer becomes increasingly malignant, abnormal cell cycle states may augment replication stress, increasing the availability of single-stranded DNA. Hence, an excess of APOBEC mutagenesis may become detectable in metastatic cancers and in later stages of tumour evolution[50,76–78] not because APOBEC is awry, but because there is more substrate for it to act upon — reflecting cumulatively disordered physiology. APOBEC mutagenesis could thus be a secondary phenomenon and importantly, under particular circumstances, could serve as a marker of tumours with highly dysregulated cell cycle control. Note that APOBEC 'drivers' or 'amplifications' have never been reported, and thus APOBEC mutagenesis does not possess the characteristics usually associated with an activating oncogenic process. Instead, it has been associated with germ line predisposition alleles that carry modest effect sizes[135–137], and is seen in many cell types, including normal cells[23,66,67,138]. All these findings point to it being a common, multifactorial mutational process.

APOBEC signatures also augured adverse prognosis more directly in multiple myeloma[139]. A link between APOBEC hypermutation and dysregulation of Maf family transcription factors, including MAF and MAFB, has been proposed. These markers of poor prognosis have binding sites within the promoter regions of *APOBEC3A* and *APOBEC3B* (REF.[139]). Overexpression of *MAF* or *MAFB* is mediated by multiple myeloma translocations t(14;16) and t(14;20)[139], providing a plausible link between the translocations and an increase in mutation load and mutation type. In non-small-cell lung cancer, APOBEC signatures were enriched in patients with durable clinical benefit following pembrolizumab treatment, and mutation counts at APOBEC-specific motifs also performed better than overall mutation burden in prognostication of patients with non-small-cell lung cancer[140].

Kataegis (a form of localized hypermutation) is also caused by APOBEC and colocalizes remarkably with structural variations[1]. Mechanistically, deamination is believed to arise because of end resection at double-strand breaks, leaving exposed single-stranded DNA. Kataegis was shown to correlate with expression of PD1 ligands PDL1 and PDL2 (REF.[141]), which are candidate biomarkers for response to immunotherapy[142]. Nonetheless, kataegis is relatively non-specific, can be present as small streaks in defined regions or can be present in large quantities around known driver amplicons[1,24]. It therefore remains unclear how best to utilize this signature as a biomarker.

In all, APOBEC-related signatures hold promise as potential biomarkers of sensitivity to immunotherapies[140] and/or as prognosticators of poor outcomes[76,143]. However, further studies are required to distinguish APOBEC mutagenesis that is truly clinically meaningful from that which is not threatening as these signatures are widespread and seen in healthy, normal cells as well[23,67].

Kataegis
A base substitution hypermutation that comprises C · G → T · A transitions and C · G → G · C transversions with a predilection for a thymine preceding the mutated cytosine (that is, a TpC context); it usually macroscopically colocalizes with structural variation.

Tandem duplicator phenotypes. Studies have reported a spectrum of TD signatures, with distinctive size variations and different driver associations indicative of diverse biological states[4,5,24,30,144,145]. In a reconstituted Tus-*Ter* system in mammalian cells, BRCA1 suppressed the formation of short TDs at site-specific replication fork barriers and not at generic double-strand breaks, indicating that TDs formed at stalled replication forks[146]. Posited to be due to a replication restart–bypass mechanism or microhomology-mediated break-induced replication[146] (FIG. 2b), the BRCA1-TD phenotype was also one of the features used in HRDetect to identify homologous recombination-deficient tumours[104].

Notably, a phenotype of longer TDs (more than 10 kb) not associated with BRCA1 loss has also been reported[4,24,145]. In a recent large-scale exercise involving more than 3,000 WGS primary cancers, TDs could be separated into two distinctive phenotypes: one characterized by TDs of 10 kb to 1 Mb in length (RS1) and another characterized by extremely long TDs (more than 1 Mb; RS14)[4]. RS1 was associated with mutations and amplifications in the cyclin E1-encoding *CCNE1*, while RS14, seen mainly in ovarian cancers, was associated with cyclin-dependent kinase 12 gene (*CDK12*) variants[5,145,147]. In liver cancer, a long TD signature (10–100 kb) was enriched in tumours with CCNA2 or CCNE1 activation[30], implicating oncogenic-induced replication stress as its inducer[148].

In terms of clinical relevance, RS1 (REFS[4,24]) is intriguing as it causes nested TDs at particular sites in the genome[149]. These sites are transcriptionally active, susceptible to DNA damage and enriched in tissue-specific superenhancers and germ line predisposition single-nucleotide polymorphisms relevant to the tumour type in question[149]. In a sinister twist, the replicative repair–bypass mechanism that generates these long TDs tends to copy a whole superenhancer or whole gene, creating mini-amplifications of sometimes well-known oncogenic drivers such as *ESR1*, *MYC* or *ZNF217* (REFS[145,149]). In other words, while the signature is initially a passenger mutational process, its tendency to create copy number gains of oncogenic loci incurs a heavy cost whereby intermediary drivers are created throughout the genome. In this respect, it is a highly malignant signature, and unsurprisingly is associated with poor outcomes[24,30,109,150].

Other genomically unstable rearrangement phenotypes. Diverse RSs, including multiple TD and deletion phenotypes, share microhomologous and/or non-templated sequences at breakpoint junctions[1,24,151]. POLQ, a key enzyme in alternative end-joining pathways, plays a key role in catalysing DNA synthesis across resected double-strand breaks when customary homology-directed repair is unavailable[28,29]. POLQ is intrinsically mutagenic, and microhomology at breakpoint junctions, whether in conjunction with deletions or templated insertions, is reported as a telltale sign of POLQ activity[27,152,153]. It is thus likely that multiple RSs are dependent on POLQ. Intriguingly, POLQ depletion makes tumour cells more sensitive to radiation, topoisomerase inhibitors and ATR inhibitors because

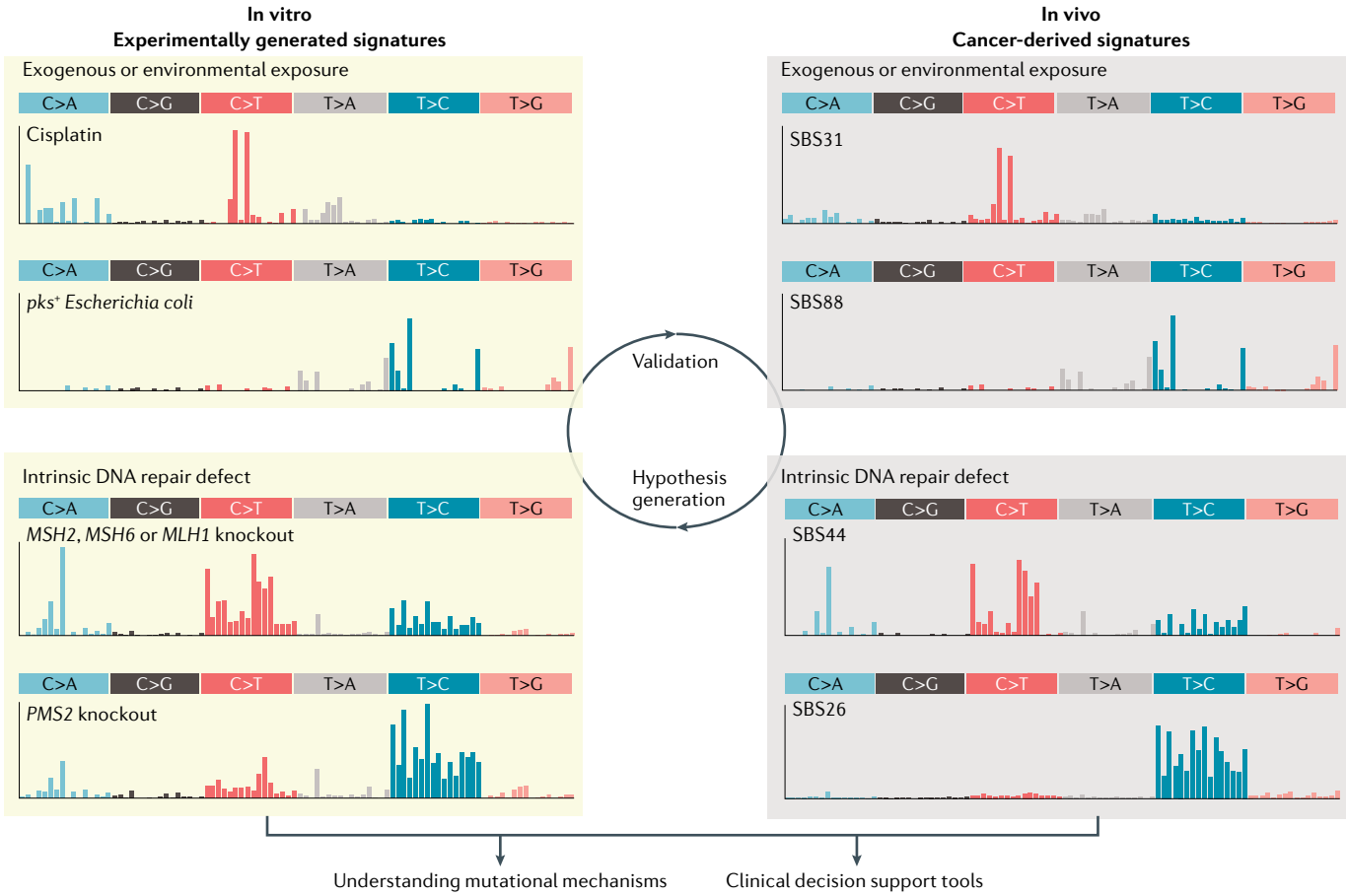


Fig. 6 | **Interplay and utility of experimental validation and cancer data analysis.** Controlled experiments based on environmental mutagen exposure and/or genetic perturbations help to elucidate the causes of mutational signatures. Associations between cancer-derived signatures and putative causes are inferred through statistical analyses of cancer data. These associations are validated through experimental studies. Both experimental and computational approaches are required to provide definitive insights into the role of mutagenesis in cancer development and support clinical algorithm development for cancer management. Some of the signature examples mentioned in the text are shown here. *pks*, polyketide synthetase genomic island, which codes for the synthesis of genotoxin colibactin; SBS, single-base substitution mutational signature. Adapted with permission from REF.[3], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/); REF.[13], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/); REF.[59], Springer Nature Limited; REF.[90], Springer Nature Limited.

of its role in repairing double-strand breaks upon fork collapse[154–156]. Therefore, these variable rearrangement phenotypes could be explored as biomarkers of tumours with synthetic lethal relationships to POLQ. In addition, particular rearrangement types, such as foldback inversions, have been shown as enriched in high-grade serous ovarian carcinoma with poor survival outcomes[157]. Finally, compound mechanisms of mutagenesis that result in various amplifications and complex chromosomal phenotypes, including extrachromosomal DNAs, are biologically interesting[158–161], and potential clinical utility may be forthcoming.

Evolvability and creation of neoepitopes. Many environmental mutagens may produce high numbers of mutations, but these mutations are often historical and do not accumulate further upon exposure cessation. By contrast, ongoing endogenous signatures that produce hypermutator phenotypes are more likely to create greater genetic diversity among daughter cells. This evolvability drives intratumour heterogeneity, increasing the likelihood of creating subclones that could develop therapeutic resistance[162]. Therefore, it is clinically important to identify ongoing signatures because they are the ones that will be targetable and/or will have predictive or prognostic value.

In line with this, ongoing mutagenesis has been reasoned to heighten sensitivity to immunotherapies[140,163]. A recent report argued that APOBEC-mediated mutations could generate neoepitopes that activate de novo T cell responses in a vaccine setting in vivo, potentially opening avenues to clinical translation[163]. In theory, if it is true that the creation of neoantigens mediates sensitivity to immunotherapy, then all the signatures that are ongoing and can create putative new antigens could be used as potential indicators of sensitivity to checkpoint blockade, including frameshifting IDs[164,165] and RSs. It remains to be seen whether this will play out.

Where to next with clinical applications. If mutational signatures are to be used in a clinically meaningful way, it will be necessary to develop assays or algorithms to

detect clinically relevant signatures robustly. With an assay in hand, it will next be imperative to demonstrate predictive or prognostic capacity through retrospective and prospective studies. This will require the collection of accompanying treatment and outcome data. An exemplar of this pathway has been shown with HRDetect[104].

Mutational signatures in liquid biopsy samples could be used to aid early cancer diagnosis, as well as in the estimation of mutagen exposure. Asymptomatic, early-stage tumours are known to leak tumour DNA into the circulation or shed circulating tumour cells into the vasculature[166,167]. Circulating tumour DNA and circulating tumour cells carry mutation information on the primary or metastatic solid tumours from which they are derived and serve as non-invasive tools for monitoring disease progression or development of resistance[168,169]. Currently, genomic analyses of liquid biopsy samples are often confined to the detection of a single or a few predefined driver mutations, mainly for surveillance[170–172]. It remains to be evaluated whether aspects of mutational signatures could be used effectively for early detection of biological states that are clinically informative[173].

Perspectives

Mutational signatures offer an additional dimension to cancer genome interpretation, a summary of biological characteristics of a tumour from each patient. To realize the full potential of whole-genome sequencing and mutational signatures, it is necessary to gain a mechanistic understanding of how mutational signatures arise through experimental exploration (FIG. 6). While we are beginning to cultivate new knowledge regarding the origins of mutational signatures, the vast majority reported to date remain of uncertain origin. Combining orthogonal modalities, including transcriptomics, proteomics and metabolomics, could offer new perspectives and reveal cooperative effects among diverse cellular processes.

A diploid whole human genome can now be sequenced at dramatically reduced costs. Continued advancements in sequencing technology, together with the development of bioinformatics tools, will truly democratize whole-genome sequencing to become an unbiased assay reporting comprehensive readouts of all clinically actionable information from a patient's tumour to make informed choices. There are currently too many trials that use single genomic points in a binary way to classify patients; for example, whether or not a tumour has a *PIK3CA* mutation in an AKT inhibitor trial. This paradigm ignores the rest of the tumour genome context. In light of recent developments and an increasing number of clinical whole-genome sequencing endeavours[174] (for example, the Beyond 1 Million Genomes (B1MG) project and the Scalable Clinical Whole-Genome Sequencing Initiative)[175], it is worth considering a shift in this ideology; to truly realize genomic potential for clinical utility, it is necessary to interpret the whole human cancer genome in its entirety, including driver mutations, germ line risk alleles and mutational signatures of all classes, to better inform how to manage each person's highly individualized cancer.

Published online 27 July 2021

1. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
This study presents catalogues of somatic mutations from 21 breast cancers, the respective mutational signatures of which were extracted by mathematical methods.
2. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
This study reports 21 distinct mutational signatures extracted from several cancer types, which form the basis of COSMIC mutational signatures v2.
3. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
This study reports the largest number of mutational signatures to date, which form the basis of COSMIC mutational signatures v3, and introduces DBSs and IDs.
4. Degasperi, A. et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).
This study introduces a practical framework and Signal, an online tool, to analyse mutational signatures. It also reports evidence of tissue-specific variability in mutational signatures, which may impact tumour classification and clinical application.
5. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
6. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
7. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
8. Secrier, M. et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).
9. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
10. Baez-Ortega, A. & Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinforma.* **20**, 77–88 (2019).
11. Omichessan, H., Severi, G. & Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS ONE* **14**, e0221235 (2019).
12. Koh, G., Zou, X. & Nik-Zainal, S. Mutational signatures: experimental design and analytical framework. *Genome Biol.* **21**, 37 (2020).
13. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
This is the largest and most comprehensive screen of environmental mutagen-associated mutational signatures published to date.
14. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
15. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
16. Fantini, D., Vidimar, V., Yu, Y., Condello, S. & Meeks, J. J. MutSignatures: an R package for extraction and analysis of cancer mutational signatures. *Sci. Rep.* **10**, 18217 (2020).
17. Cartolano, M. et al. CaMuS: simultaneous fitting and de novo imputation of cancer mutational signature. *Sci. Rep.* **10**, 19316 (2020).
18. 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).
This article describes the use of non-negative matrix factorization to extract mutational signatures.
19. Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330–337 (2018).
20. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
21. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
22. Davies, H. et al. Whole-genome sequencing reveals breast cancers with mismatch repair deficiency. *Cancer Res.* **77**, 4755–4762 (2017).
23. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
24. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
This study presents the first RSs and introduces a framework to classify these.
25. Letouze, E. et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
26. Hillman, R. T., Chisholm, G. B., Lu, K. H. & Futreal, P. A. Genomic rearrangement signatures and clinical outcomes in high-grade serous ovarian cancer. *J. Natl Cancer Inst.* **110**, 265–272 (2018).
27. Kamp, J. A., van Schendel, R., Dilweg, I. W. & Tijsterman, M. BRCA1-associated structural variations are a consequence of polymerase theta-mediated end-joining. *Nat. Commun.* **11**, 3615 (2020).
28. Mateos-Gomez, P. A. et al. Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination. *Nature* **518**, 254–257 (2015).
29. Ceccaldi, R. et al. Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. *Nature* **518**, 258–262 (2015).
30. Bayard, Q. et al. Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. *Nat. Commun.* **9**, 5235 (2018).

31. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
32. Wang, S. et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet.* **17**, e1009557 (2021).
33. Steele, C. D. et al. Undifferentiated sarcomas develop through distinct evolutionary pathways. *Cancer Cell* **35**, 441–456 (2019).
34. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
35. Lindahl, T. An *N*-glycosidase from *Escherichia coli* that releases free uracil from DNA containing deaminated cytosine residues. *Proc. Natl Acad. Sci. USA* **71**, 3649–3653 (1974).
36. Krokan, H. E. & Bjoras, M. Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5**, a012583 (2013).
37. Strauss, B. S. The "A" rule revisited: polymerases as determinants of mutational specificity. *DNA Repair* **1**, 125–135 (2002).
38. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
39. 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).
40. 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).
41. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
42. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
43. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).
44. Shen, J. C., Rideout, W. M. 3rd & Jones, P. A. High frequency mutagenesis by a DNA methyltransferase. *Cell* **71**, 1073–1080 (1992).
45. Pfeifer, G. P. Mutagenesis at methylated CpG sequences. *Curr. Top. Microbiol.* **301**, 259–281 (2006).
46. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
47. Lugli, N. et al. Enhanced rate of acquisition of point mutations in mouse intestinal adenomas compared to normal tissue. *Cell Rep.* **19**, 2185–2192 (2017).
48. Dulak, A. M. et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
49. Angus, L. et al. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* **51**, 1450–1458 (2019).
50. De Mattos-Arruda, L. et al. The genomic and immune landscapes of lethal metastatic breast cancer. *Cell Rep.* **27**, 2690–2708.e10 (2019).
51. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
52. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 4571 (2019).
53. Tomkova, M. et al. Deciphering the causes of the COSMIC mutational signature 17 by combining pan-cancer data with experimental mouse models [abstract]. *Cancer Res.* **79**, 4661 (2019).
54. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
55. Dvorak, K. et al. Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. *Gut* **56**, 763–771 (2007).
56. Inoue, M. et al. Induction of chromosomal gene mutations in *Escherichia coli* by direct incorporation of oxidatively damaged nucleotides. New evaluation method for mutagenesis by damaged DNA precursors in vivo. *J. Biol. Chem.* **273**, 11069–11074 (1998).
57. Viel, A. et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
58. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
59. Zou, X. Q. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* **2**, 643–657 (2021).
60. Kuijk, E. et al. The mutational impact of culturing human pluripotent and adult stem cells. *Nat. Commun.* **11**, 2493 (2020).
61. Rouhani, F. J. et al. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. *PLoS Genet.* **12**, e1005932 (2016).
62. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
63. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
64. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
65. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
66. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
67. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
68. D'Antonio, M. et al. Insights into the mutational burden of human induced pluripotent stem cells from an integrative multi-omics approach. *Cell Rep.* **24**, 883–894 (2018).
69. Rouhani, F. J. et al. Substantial somatic genomic variation and selection for BCOR mutations in human induced pluripotent stem cells. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.04.429731 (2021).
70. Nik-Zainal, S. & Hall, B. A. Cellular survival over genomic perfection. *Science* **366**, 802–803 (2019).
71. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184.e7 (2017).
72. Naxerova, K. et al. Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017).
73. Robinson, D. R. et al. Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
74. Yaeger, R. et al. Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. *Cancer Cell* **33**, 125–136.e3 (2018).
75. Liu, D. et al. Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer. *Nat. Commun.* **8**, 2193 (2017).
76. Swanton, C., McGranahan, N., Starrett, G. J. & Harris, R. S. APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov.* **5**, 704–712 (2015).
77. Lefebvre, C. et al. Mutational profile of metastatic breast cancers: a retrospective analysis. *PLoS Med.* **13**, e1002201 (2016).
78. Pleasance, E. et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
79. Mendelaar, P. A. J. et al. Whole genome sequencing of metastatic colorectal cancer reveals prior treatment effects and specific metastasis features. *Nat. Commun.* **12**, 574 (2021).
80. Rubanova, Y. et al. Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat. Commun.* **11**, 731 (2020).
81. Riva, L. et al. The mutational signature profile of known and suspected human carcinogens in mice. *Nat. Genet.* **52**, 1189–1197 (2020).
82. Olivier, M. et al. Modelling mutational landscapes of human cancers in vitro. *Sci. Rep.* **4**, 4482 (2014).
83. Besaratinia, A. & Pfeifer, G. P. Applications of the human p53 knock-in (Hupki) mouse model for human carcinogen testing. *FASEB J.* **24**, 2612–2619 (2010).
84. Liu, Z. et al. Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proc. Natl Acad. Sci. USA* **101**, 2963–2968 (2004).
85. Szikriszt, B. et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 99 (2016).
86. 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).
87. Volkova, N. V. et al. Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.* **11**, 2169 (2020).
88. 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).
89. Pleguezuelos-Manzano, C. et al. Mutational signature in colorectal cancer caused by genotoxic pks⁺ *E. coli*. *Nature* **580**, 269–273 (2020).
90. Dziubanska-Kusibab, P. J. et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med.* **26**, 1063–1069 (2020).
91. Boot, A. et al. Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res.* **30**, 803–813 (2020).
92. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
93. Jager, M. et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019).
94. 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).
95. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
96. Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
97. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
98. Fong, P. C. et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).
99. Telli, M. L. et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).
100. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
101. Birkbak, N. J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
102. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
103. Timms, K. M. et al. Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **16**, 475–483 (2014).
104. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017). **This study describes the first clinical predictive tool, HRDetect, designed using a panel of mutational signatures to predict HRD.**
105. Zhao, E. Y. et al. Homologous recombination deficiency and platinum-based therapy outcomes in advanced breast cancer. *Clin. Cancer Res.* **23**, 7521–7530 (2017).
106. Nones, K. et al. Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. *Ann. Oncol.* **30**, 1071–1079 (2019).
107. Nguyen, L., Martens, J. W. M., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 5584 (2020).
108. Gulhan, D. C., Lee, J. J., Melloni, G. E. M., Cortes-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).
109. Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
110. Chopra, N. et al. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. *Nat. Commun.* **11**, 2662 (2020).
111. Thibodeau, S. N., Bren, G. & Schaid, D. Microsatellite instability in cancer of the proximal colon. *Science* **260**, 816–819 (1993).

112. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**, 558–561 (1993).
113. Kim, T. M., Laird, P. W. & Park, P. J. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155**, 858–868 (2013).
114. Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895–2015. *Nat. Rev. Cancer* **15**, 181–194 (2015).
115. Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
116. Mandal, R. et al. Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. *Science* **364**, 485–491 (2019).
117. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
118. Middha, S. et al. Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. *JCO Precis. Oncol.* **1**, 1–17 (2017).
119. Niu, B. F. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
120. Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. *Clin. Chem.* **60**, 1192–1199 (2014).
121. Germano, G. et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116–120 (2017).
122. Lemery, S., Keegan, P. & Pazdur, R. First FDA approval agnostic of cancer site — when a biomarker defines the indication. *N. Engl. J. Med.* **377**, 1409–1412 (2017).
123. Stelloo, E. et al. Practical guidance for mismatch repair-deficiency testing in endometrial cancer. *Ann. Oncol.* **28**, 96–102 (2017).
124. Kawakami, H., Zaanan, A. & Sinicrope, F. A. Microsatellite instability testing and its role in the management of colorectal cancer. *Curr. Treat. Options Oncol.* **16**, 30 (2015).
125. Buhard, O. et al. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. *J. Clin. Oncol.* **24**, 241–251 (2006).
126. Huang, M. N. et al. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci. Rep.* **5**, 13321 (2015).
127. Fabrizio, D. A. et al. Beyond microsatellite testing: assessment of tumor mutational burden identifies subsets of colorectal cancer who may respond to immune checkpoint inhibition. *J. Gastrointest. Oncol.* **9**, 610–617 (2018).
128. Schrock, A. B. et al. Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. *Ann. Oncol.* **30**, 1096–1103 (2019).
129. Mehnert, J. M. et al. Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. *J. Clin. Invest.* **126**, 2334–2340 (2016).
130. Howitt, B. E. et al. Association of polymerase e-mutated and microsatellite-instable endometrial cancers with neoantigen load, number of tumor-infiltrating lymphocytes, and expression of PD-1 and PD-L1. *JAMA Oncol.* **1**, 1319–1323 (2015).
131. Johanns, T. M. et al. Immunogenomics of hypermutated glioblastoma: a patient with germline POLE deficiency treated with checkpoint blockade immunotherapy. *Cancer Discov.* **6**, 1230–1236 (2016).
132. Momen, S. et al. Dramatic response of metastatic cutaneous angiosarcoma to an immune checkpoint inhibitor in a patient with xeroderma pigmentosum: whole-genome sequencing aids treatment decision in end-stage disease. *Cold Spring Harb. Mol. Case Stud.* **5**, a004408 (2019).
133. Chung, J. et al. DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. *Cancer Discov.* **11**, 1176–1191 (2020).
134. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
135. Nik-Zainal, S. et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487–491 (2014).
136. Starrett, G. J. et al. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. *Nat. Commun.* **7**, 12918 (2016).
137. Middlebrooks, C. D. et al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* **48**, 1330–1338 (2016).
138. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
139. Walker, B. A. et al. APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat. Commun.* **6**, 6997 (2015).
140. Wang, S. X., Jia, M. M., He, Z. K. & Liu, X. S. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene* **37**, 3924–3936 (2018).
141. Boichard, A., Tsigelny, I. F. & Kurzrock, R. High expression of PD-1 ligands is associated with kataegis mutational signature and APOBEC3 alterations. *Oncoimmunology* **6**, e1284719 (2017).
142. Gibney, G. T., Weiner, L. M. & Atkins, M. B. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol.* **17**, e542–e551 (2016).
143. Law, E. K. et al. The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Sci. Adv.* **2**, e1601737 (2016).
144. Menghi, F. et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl Acad. Sci. USA* **113**, E2373–E2382 (2016).
145. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210.e5 (2018).
146. Willis, N. A. et al. Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature* **551**, 590–595 (2017).
147. Popova, T. et al. Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. *Cancer Res.* **76**, 1882–1891 (2016).
148. Macheret, M. & Halazonetis, T. D. Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. *Nature* **555**, 112–116 (2018).
149. Glodzik, D. et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat. Genet.* **49**, 341–348 (2017).
150. Quigley, D. A. et al. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **174**, 758–769.e9 (2018).
151. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
152. Schimmel, J., Kool, H., van Schendel, R. & Tijsterman, M. Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. *EMBO J.* **36**, 3634–3649 (2017).
153. Wyatt, D. W. et al. Essential roles for polymerase theta-mediated end joining in the repair of chromosome breaks. *Mol. Cell* **63**, 662–673 (2016).
154. Higgins, G. S. et al. A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. *Cancer Res.* **70**, 2984–2993 (2010).
155. Yousefzadeh, M. J. et al. Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. *PLoS Genet.* **10**, e1004654 (2014).
156. Wang, Z. et al. DNA polymerase (POLQ) is important for repair of DNA double-strand breaks caused by fork collapse. *J. Biol. Chem.* **294**, 3909–3919 (2019).
157. Wang, Y. K. et al. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nat. Genet.* **49**, 856–865 (2017).
158. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163**, 1641–1654 (2015).
159. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
160. Umbreit, N. T. et al. Mechanisms generating cancer genome complexity from a single cell division error. *Science* **368**, 282–294 (2020).
161. Shoshani, O. et al. Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**, 137–141 (2021).
162. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
163. Driscoll, C. B. et al. APOBEC3B-mediated corruption of the tumor cell immunopeptidome induces heteroclitic neoepitopes for cancer immunotherapy. *Nat. Commun.* **11**, 790 (2020).
164. Roudko, V. et al. Shared immunogenic poly-epitope frameshift mutations in microsatellite unstable tumors. *Cell* **183**, 1634–1649.e17 (2020).
165. Koster, J. & Plasterk, R. H. A. A library of neo open reading frame peptides (NOPs) as a sustainable resource of common neoantigens in up to 50% of cancer patients. *Sci. Rep.* **9**, 6577 (2019).
166. Diaz, L. A. Jr & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* **32**, 579–586 (2014).
167. Alix-Panabieres, C. & Pantel, K. Clinical applications of circulating tumor cells and circulating tumor dna as liquid biopsy. *Cancer Discov.* **6**, 479–491 (2016).
168. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
169. Annala, M. et al. Circulating tumor DNA genomics correlate with resistance to abiraterone and enzalutamide in prostate cancer. *Cancer Discov.* **8**, 444–457 (2018).
170. Dawson, S. J. et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
171. Murtaza, M. et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
172. Misale, S. et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532–536 (2012).
173. Zviran, A. et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* **26**, 1114–1124 (2020).
174. Turnbull, C. et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).
175. Weill Cornell Medicine. Weill Cornell Medicine, NewYork-Presbyterian Hospital, and Illumina collaborate on scalable clinical whole-genome sequencing initiative. *EurekAlert* https://www.eurekalert.org/pub_releases/2020-12/wcm-wcm120220.php (2020).
176. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).

Acknowledgements
This work was funded by the Cancer Research UK (CRUK) Advanced Clinician Scientist Award (C60100/A23916), the Dr. Josef Steiner Cancer Research Award 2019, a Medical Research Council (MRC) Grant-in-Aid to the MRC Cancer Unit, the CRUK Pioneer Award, a Wellcome Strategic Award (WT101126), Basser Gray Prime Award 2020, and supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014).

Author contributions
S.N.-Z., G.K., A.D. and S.M. researched data for the article. S.N.-Z., G.K., A.D. and X.Z. contributed to discussion of content and writing the article. S.N.-Z., G.K. and X.Z. reviewed and edited the manuscript before submission.

Competing interests
S.N.-Z. holds patents on clinical algorithms of mutational signatures and, during the completion of the manuscript, also had advisory roles for AstraZeneca, Artios Pharma Ltd and the Scottish Genome Project. The other authors declare no competing interests.

Peer review information
Nature Reviews Cancer thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41568-021-00377-7.

RELATED LINKS
Beyond 1 Million Genomes (B1MG): https://b1mg-project.eu
Signal: a Web-based tool for cancer and experimentally generated mutational signature exploration and analysis: https://signal.mutationalsignatures.com
Wellcome Trust Sanger Institute COSMIC signature resource: https://cancer.sanger.ac.uk/cosmic/signatures