Journal of Pathology
J Pathol July 2022; **257:** 454–465
Published online 20 May 2022 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/path.5912

INVITED REVIEW

An overview of mutational and copy number signatures in human cancer

Christopher D Steele[1†*], Nischalan Pillay[1,2†] and Ludmil B Alexandrov[3,4,5†]

[1] *Research Department of Pathology, Cancer Institute, University College London, London, UK*
[2] *Department of Cellular and Molecular Pathology, Royal National Orthopaedic Hospital NHS Trust, Stanmore, UK*
[3] *Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, USA*
[4] *Department of Bioengineering, UC San Diego, La Jolla, CA, USA*
[5] *Moores Cancer Center, UC San Diego, La Jolla, CA, USA*

**Correspondence to: CD Steele, Research Department of Pathology, Cancer Institute, University College London, London, WC1E 6BT, UK. E-mail: c.steele.11@ucl.ac.uk*

†The authors contributed equally to this work.

Abstract

The genome of each cell in the human body is constantly under assault from a plethora of exogenous and endogenous processes that can damage DNA. If not successfully repaired, DNA damage generally becomes permanently imprinted in cells, and all their progenies, as somatic mutations. In most cases, the patterns of these somatic mutations contain the tell-tale signs of the mutagenic processes that have imprinted and are termed mutational signatures. Recent pan-cancer genomic analyses have elucidated the compendium of mutational signatures for all types of small mutational events, including (1) single base substitutions, (2) doublet base substitutions, and (3) small insertions/deletions. In contrast to small mutational events, where, in most cases, DNA damage is a prerequisite, aneuploidy, which refers to the abnormal number of chromosomes in a cell, usually develops from mistakes during DNA replication. Such mistakes include DNA replication stress, mitotic errors caused by faulty microtubule dynamics, or cohesion defects that contribute to chromosomal breakage and can lead to copy number (CN) alterations (CNAs) or even to structural rearrangements. These aberrations also leave behind genomic scars which can be inferred from sequencing as CN signatures and rearrangement signatures. The analyses of mutational signatures of small mutational events have been extensively reviewed, so we will not comprehensively re-examine them here. Rather, our focus will be on summarising the existing knowledge for mutational signatures of CNAs. As studying CN signatures is an emerging field, we briefly summarise the utility that mutational signatures of small mutational events have provided in basic science, cancer treatment, and cancer prevention, and we emphasise the future role that CN signatures may play in each of these fields.

Keywords: mutational signatures; copy number signatures; genomics; cancer

Received 6 March 2022; Revised 8 April 2022; Accepted 11 April 2022

Conflict of interest statement: NP is an Associate Editor of The Journal of Pathology *and a member of the Education Committee of The Pathological Society of Great Britain and Ireland (owners of the journal). LBA is an inventor of a US Patent (10,776,718) for source identification by non-negative matrix factorisation.*

The difference between DNA damage and somatic mutations

Some processes, mainly those due to internal cellular mechanisms, are ubiquitous and can be found in every cell. Examples of such ubiquitous processes include generation of reactive oxygen species, as part of the normal functioning of the mitochondria, or internal cellular processes inadvertently mutating DNA during replication [1]. Other processes are tissue-specific or can be due to lifestyle choices or to environmental exposures. For example, ultraviolet radiation (UV) in sunlight will damage the DNA of skin cells, but it will not affect cells in internal organs (e.g. cells in the liver or pancreas). Similarly, smoking tobacco cigarettes will damage the DNA of certain organs, most prominently cells of the lung, but it will not alter the genomes of cells in other organs, such as the brain. These endogenous and exogenous processes result in DNA damage. In general, DNA damage refers to *chemical modifications* of nucleotides (adenine, thymine, guanine, and cytosine) as well as to changes/breaks in the covalent bonds between adjacent nucleotides or the hydrogen bonds between traditional DNA base-pairs. Some examples of DNA damage are oxidation, deamination, and alkylation of DNA bases; formation of single- and double-strand breaks; formation of covalent linkages between consecutive bases along the nucleotide chain; and many others [2]. In contrast, somatic mutations reflect a *change in the sequence* of the base-pair content in DNA. Somatic mutations could include small mutational events such as substitutions, where a DNA base-pair is substituted with another DNA base-pair (e.g. C:G base-pair changes to a T:A base-pair), and small insertions/deletions, where, for example, a set of DNA base-pairs is completely removed from the genome (e.g. a set of CAC:GTG base-pairs gets deleted from the genome). Somatic mutations can also include large mutational events affecting many millions of base pairs across multiple chromosomes. Examples of such types of mutations generally include genomic structural variants (SVs), where, for example, two chromosomes can become partially fused, or copy number (CN) variations, where the number of copies of a large genomic segment gets amplified or deleted. Some examples of CN changes are a loss of a chromosome arm, doubling the number of copies of a specific chromosome, and even the entire diploid genome: doubling the number of copies (whole-genome duplication) or losing a copy of the diploid genome (haploidisation).

An important difference between *damaged* and *mutated* DNA is that while some DNA damage may persist for decades, it is generally confined to the cell in which it occurred [3]. In contrast, somatic mutations affect both their cell of origin and all progenitors of that cell, thus having an exponential functional effect in all offspring. While, in many cases, somatic mutations require pre-existing DNA damage, the road from a damaged DNA to a somatic mutation involves multiple additional molecular steps – with the vast majority of DNA damage being repaired and not resulting in any somatic mutations. However, a failure of a DNA repair pathway in a cell can result in a specific type of DNA damage being un- or under-repaired, causing a mutator phenotype, and resulting in an explosion of somatic mutations, e.g. mismatch repair deficiency [4]. Importantly, not all mutations necessitate prior DNA damage; examples include infidelity of the polymerases during replication, leading to substitutions or indels in progenitor cells, as well as large CN events such whole-genome doubling caused through mitotic error [5].

In most cases, sequencing the genomes of cancer or normal somatic cells allows the derivation of somatic mutations, but it does not directly allow the study of DNA damage. Nevertheless, as the activity of DNA damage and repair processes reflects many of the imprinted mutations, analysis of somatic mutations may also be leveraged to understand DNA damage and repair processes. Different types of sequencing approaches provide different resolutions in examining different types of somatic mutations [6]. A complete sequencing of a genome (also known as whole-genome sequencing, WGS) allows examination for all types of somatic mutations, which include (1) single base substitutions (SBSs); (2) small indels and deletions (indels); (3) copy number alterations (CNAs); and (4) genomic SVs. In contrast, sequencing of all exons of the genome, also known as whole-exome sequencing (WES), allows interrogation for substitutions and indels in coding regions as well as providing limited resolution of genome-wide CNAs. Sequencing only a small set of actionable genes, such the targeted panels commonly used for cancer diagnosis, can be utilised to detect substitutions and indels in these genes and may provide an extremely low-resolution map for detecting certain very large genomic segments with CNAs. In most cases, whole-exome and targeted-panel sequencing do not allow detection of SVs, unless the experimental protocols have been designed to detect specific SVs.

Understanding somatic mutagenesis through mutational signatures

Somatic mutations have been shown to play a pivotal role in cancer initiation and cancer evolution [7,8]. Mutations that impact gene functionality, either through a gain or through a loss of function, and enhance the fitness characteristics of a cell are often referred to as *driver mutations*. While non-neoplastic and cancer cells harbour fewer than ten driver mutations, their genomes are moulded with many thousands (and, in some cases, even millions) of somatic mutations that provide little fitness advantage; such mutations are generally termed *passenger mutations* [9]. The distribution of passenger mutations across the genomic landscape can be affected by several different factors that alter DNA damage, DNA repair, or other cellular processes. From a genomic architecture perspective, the locations of passenger mutations are modified by replication timing, transcriptional activity, eu- and hetero-chromatin, histone modifications, transcription factor binding sites, and factors related to nucleosome occupancy [10]. From a local sequencing context perspective, different mutational processes have different biophysical and biochemical characteristics, resulting in a specific preference for the immediate sequence context of the imprinted somatic mutations. This combination of micro- and macro-genomic features affects the accumulation of somatic mutations from a particular mutational process, leading to a characteristic pattern of somatic mutations, termed the *mutational signature*.

The first descriptions of specific patterns of DNA damage identified through exogenous sources were presented in the late 1950s, when it was discovered that UV light exposure of *in vitro* systems resulted in the formation of pyrimidine photodimers (two consecutive bases on one strand that bind together), predominantly affecting adjacent cytosine and thymine nucleotides and leading to cytosine–cytosine, cytosine–thymine, thymine–cytosine, and thymine–thymine photodimers [11–13]. Subsequent work described the biophysical and biochemical properties leading to the formation of DNA damage from UV light, including both cyclobutane pyrimidine dimers (CPDs) and pyrimidine(6-4)pyrimidone photoproducts.

Almost 20 years later, with the advent of Sanger sequencing, a causative molecular link was made between UV-induced DNA damage, subsequent formation of somatic mutations, and the development of skin cancer by examining the patterns of mutations in certain skin cancers [14]. Similar research was also performed for a number of other cancer types by evaluating the patterns of mutations in the exons of *TP53*, the most commonly mutated gene in human cancer, revealing a number of distinct mutational patterns including ones attributed to endogenous deamination, tobacco smoking, aflatoxin, and others.

The advances in next-generation sequencing technologies have facilitated the unbiased assessment of the mutational patterns of cancer genomes [15,16]. By sequencing the genome of a cancer, one can now observe the combined outcome of all mutational forces that have been active at different strengths and at different times throughout the lineage of the cancer cell. Importantly, by utilising a proper mathematical model and a set of machine learning (ML) computational approaches, one can decipher the individual signature of each process that has been active throughout this lineage. Moreover, by uniting the ML-derived mutational signatures with experimental data, one can even understand the mutational processes that gave rise to these signatures and that caused the mutations in a cancer patient [17].

During the last decade, analysis of mutational signatures has become a standard approach in examining somatic mutations derived from next-generation sequencing data (see [2,18,19] for in depth reviews of mutational signatures). In the majority of cases, these analyses have utilised somatic mutations derived from whole-exome and/or whole-genome sequenced cancer genomes by applying unsupervised ML approaches based on non-negative matrix factorisation (NMF). In brief, from an analysis perspective, the mutations in a set of cancer genomes are categorised based on a mutational classification into distinct categories resulting into a mutational matrix, where each column reflects a cancer genome, each row a distinct mutation type, and the value of each cell corresponds to the number of mutations of a particular mutation type in a cancer genome. The mutational matrix is subsequently factorised with NMF into two matrices, one reflecting the mutational signatures and the other corresponding to the activities of each signature in each sample. Initially, analysis of mutational signatures across cancer types was performed only for SBSs, due to the simplicity in classifying these types of mutations, but has now expanded to include SBSs, doublet base substitutions (DBSs), small insertions and deletions (IDs), SVs, and CN signatures [20–24].

Biological underpinnings and clinical implications of mutational signatures

From a research perspective, mutational signatures of small mutational events have provided a novel toolset for indirectly studying the molecular processes of DNA damage, DNA repair, and even DNA replication. Analyses of cancer genomes have elucidated the effect of genome architecture and the topographical features of the human genome on the cancer-specific accumulation of somatic mutations from some mutational signatures [19]. Experimental works have also revealed the compendium of *in vitro*-induced mutational signatures and the vulnerabilities of human stem cells to different endogenous and exogenous mutagens. Importantly, examination of mutational signatures has brought significant insights into the interactions between mismatch repair and replication; enzymatic deamination by AID/APOBEC3; transcription-coupled repair; transcription-coupled damage; topoisomerases and DNA repair; clustered mutagenesis and genomic rearrangements; and many others [25–31].

From a cancer prevention perspective, analysis of SBS, DBS, and ID mutational signatures has revealed a number of environmental mutagens causing specific cancer types with notable examples including (1) aristolochic acid: a group of acids found naturally in many types of plants, which has been linked to cancers of the liver, bladder, kidney, oral cavities, and oesophagus; (2) aflatoxin: a family of toxins produced by certain fungi, linked to liver cancer; (3) colibactin: a potent genotoxin associated with certain strains of *Escherichia coli*, found in colorectal cancer; and many others [32–35]. These findings have allowed the proposal and development of strategies for preventing cancer by limiting exposure to such mutagens. Additionally, mutational signatures have been used as multi-tumour phenotypes of germline predisposition, which has allowed for better screening of people with a higher risk for developing cancer. A notable example is that of *NTHL1*, where germline mutations can give rise to a multitude of different cancer types, which, prior to mutational signatures, was not fully appreciated [36].

From a cancer treatment perspective, mutational signatures have proven to be a valuable resource both for understanding iatrogenic exposures leading to secondary/recurrent cancers and for optimally targeting cancer therapies. Several studies have revealed the mutational signatures of chemo- and immune-therapies, demonstrating, in many cases, that secondary cancers are caused by specific treatment regimens. Notable examples include secondary cancers after treatment with temozolomide and platinum therapies, as well as the role of azathioprine – a drug used in auto-immune conditions – in causing primary cancers. Importantly, mutational signatures have also been shown to have clinical utility in selecting an optimum treatment strategy. A presence of specific mutational signatures has been used as a predictive biomarker for response to PARP inhibitors, platinum therapy, immunotherapy, and tamoxifen resistance [37–39].

Large-scale national and international cancer genome sequencing efforts, such as The Cancer Genome Atlas (TCGA) project and the International Cancer Genome Consortium (ICGC), have profiled many thousands of human cancers, thereby producing a rich resource consisting of terabytes of somatic mutational data [40,41]. This has catalysed a new field of research which has now been extended to the analysis of CN and SV signatures.

The lexicon of copy number profiling: the devil is in the detail

The evaluation of chromosomes in cells through karyotyping is an established technique to evaluate aneuploidy either in the germline setting, e.g. to identify trisomy of chromosome 21 in Down's syndrome, or in the somatic setting, e.g. to reveal chromosomal changes in cancer [42]. This laboratory technique provides a global low-resolution snapshot of CN aberrations that include large gains and losses of DNA, duplicated chromosomes, and translocations [43]. Higher throughput methods using comparative genomic hybridisation (CGH) arrays or single nucleotide polymorphism (SNP) arrays with vastly increased resolution allow for genome-wide interrogation of CN down to the gene level on fresh, frozen, and even formalin-fixed, paraffin-embedded tissues, which has allowed for unprecedented insights into the CN landscape of cancer [44,45]. The advantage of SNP arrays over CGH is that the genotyping of common SNPs enables the ability to infer not only the total CN by using the signal intensity (i.e. log*R*) across a region of the genome but also the ratio of reference and alternate allele intensities (B-allele frequency) to elucidate the allele-specific CN changes within the same region [46]. Next-generation sequencing technologies, such as WES and WGS, can also be utilised to generate allele-specific CN across the genome [47], by deriving log*R* from the sequencing depth and the B-allele frequencies. This combination of log*R* and B-allele frequency for a given sample is commonly referred to as the *copy number profile* of the sample (Figure 1). Multiple bioinformatic tools to derive CN profiles have been developed, ranging from relatively straightforward methods that utilise log*R* values to infer total CN [48], to methods that use both log*R* and BAF values to infer allele-specific CN [46,49–51], and even more sophisticated methods that infer not only clonal CN but also sub-clonal CNAs through the application of haplotype phasing [25]. In addition, the choice of CN caller is often motivated by the utilised profiling platform, e.g. B-allele frequencies are challenging to obtain from shallow WGS data (low sequencing coverage WGS), whereas haplotype phasing is challenging for non-WGS data. Recent advances that combine information across multiple samples have improved CN profiling where samples are evolutionarily related, e.g. multi-region sampling of a single tumour [52], while also paving the way for allele-specific CN calling from shallow WGS and from single-cell sequencing data [53].

For CN analysis, probably the most commonly used approach to decipher potential CN drivers from cancer genome samples is the Genomic Identification of Significant Targets in Cancer (GISTIC) tool [54]. Using this method, one can identify statistically significant regions of recurrent amplifications and deletions, in essence providing a method to identify potential CN driver mutations. Whilst extremely useful, the GISTIC method does not allow one to infer the mutational processes that may be generating those CN changes. In an analogous way to mutational signatures, the newly developed CN signature methods relate CN patterns to potential mutational processes.

A deep look into the short history of copy number signatures

The field of mutational signature analysis has developed multiple methods based on non-negative matrix factorisation, latent Dirichlet allocation, hierarchical Dirichlet processes, or other methods for data classification

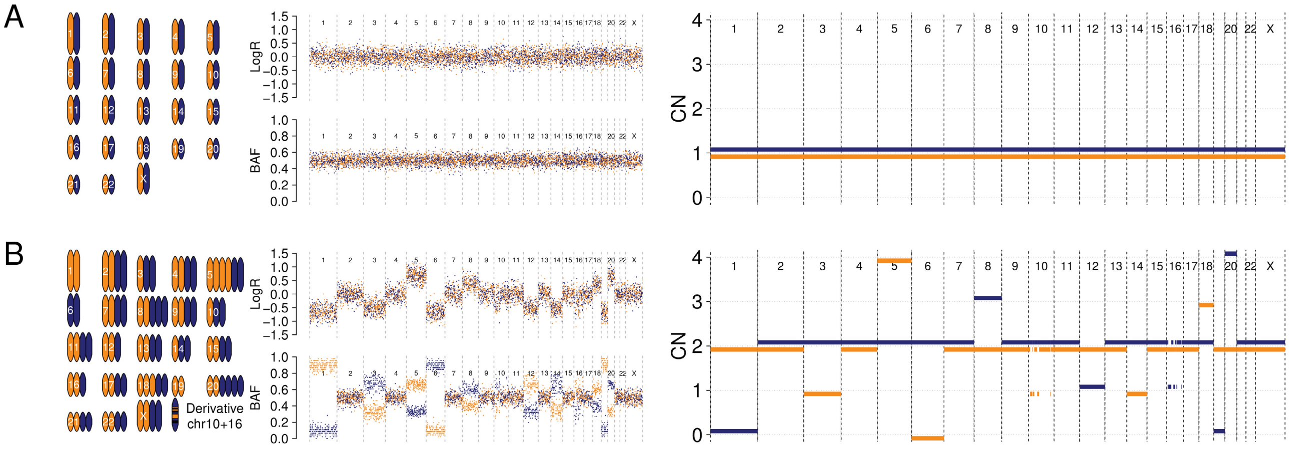

Figure 1. Examples of CN profiles derived from karyotyping and DNA sequencing. Mock karyotype (left), log*R* and BAF tracks (middle), and CN profile (right) for (A) a diploid and (B) an aneuploid genome. Maternal (blue) and paternal (orange) chromosomes are displayed separately and phased throughout. CN, copy number. The aneuploid genome is whole-genome doubled (WGD) and includes losses prior to (chr1 + 6 + 19) and following (chr3 + 12 + 14) WGD, as well as gains prior to (chr5 + 20) and following (chr8 + 18) WGD. Additionally, a chromothripsis event after WGD involving both chr10 and chr16 is included.

Table 1. Design of features for CN signature methods

Reference	Genomic feature	Biological process
Macintyre *et al* (2018) [22]	Breakpoint count per 10 Mb	Chromothripsis
	CN	
	CN change point	Breakage fusion bridge
	Breakpoint count per chromosome arm	Chromothripsis
	Length of chains of oscillating CN	Chromothripsis
	Segment size	HRD, chromothripsis
Wang *et al* (2021) [62]*	Minimum number of chromosomes with 50% CN altered (1 value)	Ploidy
	Number of events per chromosome	Genomic distribution of events
Steele *et al* (2019/2021) [23,24]	LOH status	Mechanism-agnostic†
	Total CN	
	Segment size	

*Additional features introduced to the method described in ref 22.

†Patterns identified from mechanism-agnostic approaches can be associated with various datasets *post hoc*; this allows for linking of multiple mechanism-agnostic signatures with processes such as genome doubling, chromosomal instability, chromothripsis, HRD, and haploidisation [24].

[15,54–56] and tools [15,57–61] to extract meaningful information from cancer genomes. An early study in 2018 on high-grade serous ovarian cancer (OV) was one of the first to use such mathematical approaches to show the potential utility of a CN signature for prognostication [22]. OV is a genetically complex tumour characterised by *TP53* mutations, often associated with homologous recombination deficiency (HRD) (see below), breakage fusion bridge cycles – a pattern of replication-associated genomic instability induced by telomere shortening, leading to amplifications and genomic rearrangements, and chromothripsis, which is a mutational process leading to clustered rearrangements that occur as a single event following the 'shattering of a chromosome' [55]. In their CN signature framework, Macintyre *et al* [22] designed a model to capture these salient genetic features of the OV genome through CN signatures derived from shallow WGS data (Table 1). Their analysis demonstrated the potential clinical utility of CN signatures by identifying a signature that was linked to poor prognosis in some patients. OV served as an ideal model upon which to build such a framework, as the endogenous DNA damage processes linked with the CN patterns were already well established in that cancer type [55,56].

The utility of signatures as biomarkers for treatment stratification in cancer is perhaps best exemplified by those developed for identifying HRD [22,24]. Inactivating mutations in *BRCA1*, *BRCA2* or other HR pathway genes leave tumour cells unable to repair double-strand breaks of the genome in a faithful manner, instead relying on error-prone pathways such as non-homologous end joining or microhomology-mediated end joining [57,58]. Antagonising this deficiency, either through inducing DNA damage (as for platinum-based therapies) or through blocking compensatory repair pathways (as for PARP inhibition), is a promising therapeutic avenue [59]. Nevertheless, stratifying patients for these treatments remains an open challenge. The US Food and Drug Administration has approved the Myriad Genetics myChoice CDx® test, which combines identification of inactivating mutations of *BRCA1* or *BRCA2* along with specific CN-based readouts of 'genomic scars' of HRD [60]. However, alterations of other HR pathway genes, or non-genetic mechanisms, such as promoter methylation silencing [61], may abrogate the HR pathway but could be missed by the myChoice CDx® test. Similar CN-based metrics inspired by the myChoice CDx® genomic scars have been developed [63]; however, the application of such metrics across different tumour types requires careful calibration [64]. Beyond CN genomic scars, signatures of single base substitutions, indels, and rearrangements have been associated with HRD tumours [15,16,21]. In an effort to generate a robust test for HRD, Davies *et al* built a predictive model for HRD – HRDetect – that incorporates SBS, indel, and rearrangement signatures, as well as CN genomic scars [65,66], which holds great promise but is restricted to WGS data. Other methods that are dependent on whole-genome sequencing [67] or that have removed the need for WGS data and are therefore applicable to targeted sequencing panels [68] have also been published. Additionally, the distinction between ongoing HRD and historic HRD needs to be properly assessed in tumours; historic HRD may leave the scars on the genome that indicate that the tumour is HR-deficient, but the tumour may have reinstated HR through mechanisms such as *PTEN* mutation, or compensating mutations up- or down-stream of the core HR pathway genes [69]. This distinction is important for patients, as ongoing HRD should be sensitive to PARP inhibition, whereas tumours that have reinstated HR may be resistant.

In 2019, through a separate study of genomically complex undifferentiated soft tissue sarcomas (USARCs), where limited prior knowledge about the underlying mutational processes was known, Steele *et al* developed a 'mechanism-agnostic' approach to summarise CN profiles into CN signatures [23]. This was achieved by categorising the segments of an allele-specific CN profile by fundamental aspects of CN into a CN summary vector, which included loss of heterozygosity (LOH) status, total CN state, and segment size (Figure 2A,B). Once a set of CN summaries is obtained (Figure 3A), the matrix of all summaries can be decomposed into a matrix of signature definitions and a matrix of signature attributions (Figure 3B) using non-negative matrix factorisation. The set of signatures and their attributions can then be used to infer the potential biological relevance of the signatures in question

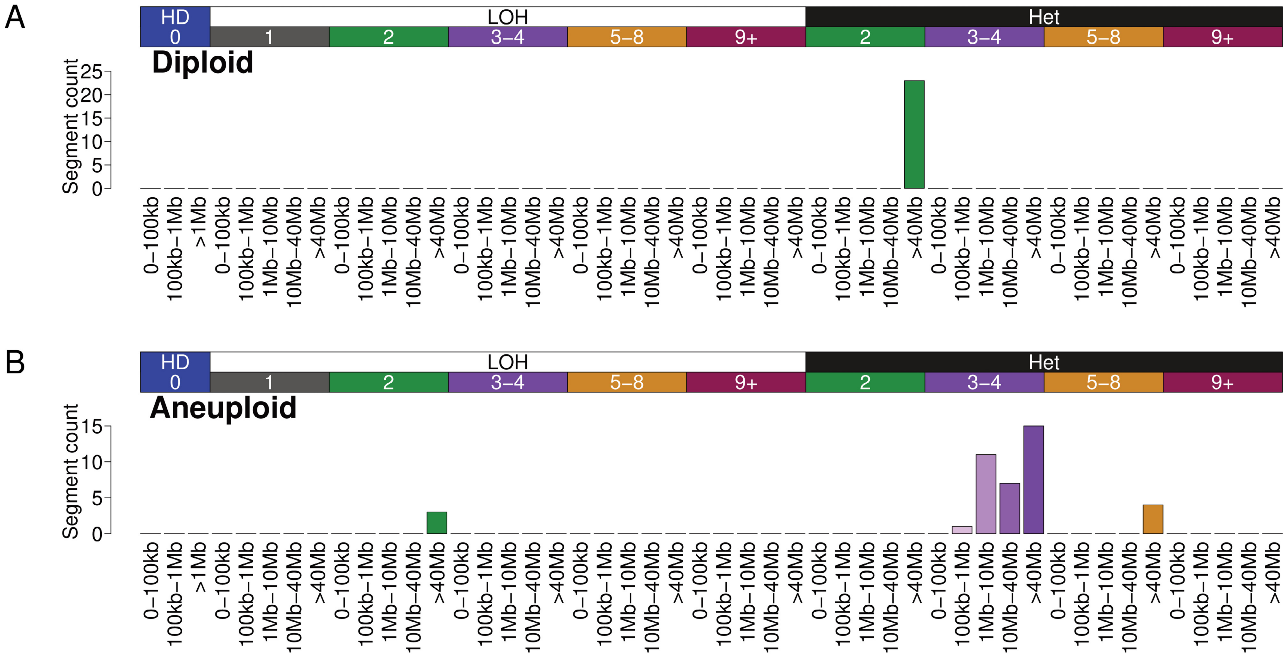

Figure 2. CN summary vectors for the two profiles displayed in Figure 1. (A) Diploid and (B) aneuploid, as per the summary methodology of Steele *et al* [24]. LOH status and total CN categories are displayed above the bar plots. Segment size categories are displayed below the bar charts.

through associations with relevant data such as driver mutations, chromothripsis, whole-genome doubled (WGD) status, or even orthogonal information such as transcriptome profiling [22–24,62,70]. In USARCs, seven distinctive signatures were identified which were linked to biological processes including successive whole-genome doublings, genome-wide LOH events, and chromothripsis [23]. We were able to integrate these signatures with other genomic data to elucidate multiple evolutionary routes through which USARCs potentially develop, demonstrating the utility of CN signatures for understanding the evolutionary history of cancer genomes. Recently, Steele *et al* expanded this work to ~10 000 cancers to develop a pan-cancer set of 21 CN signatures [24], with new signatures linked to HRD, extrachromosomal circular DNA formation, and haploidisation. Moreover, this work provided a further refinement in the evolutionary relationship between CN signatures and their role in tumour history.

The two signature methods described above broadly divide the field into two major classes: those that have CN features designed specifically to capture previously known biological processes [22,62,70] and those designed to be mechanism-agnostic [23,24] (Table 1). Importantly, patterns identified from mechanism-agnostic approaches can be subsequently associated with various datasets *post hoc*; this has allowed the linking of multiple mechanism-agnostic signatures with processes such as genome doubling, chromosomal instability, chromothripsis, HRD, and haploidisation [24]. While mechanism-agnostic approaches may not capture all known biological processes, they can reveal previously unappreciated molecular mechanisms [24]. In contrast, while CN features designed to capture a particular set of biological processes will capture these processes, they would rarely elucidate previously unknown molecular mechanisms.

Macintyre *et al* [22] designed a set of CN features that would each tag a given process known to generate CN profiles with distinctive values for those features, e.g. classical chromothripsis will generate many breakpoints with long oscillating CN runs of short segment size [71,72] (Table 2). In addition to the studies outlined above, a set of other CN signature analyses were also performed, with most of them focusing on detecting specific biological processes. For example, the above set of features was expanded by Wang *et al* [62] to include metrics of global CN aberration extent and local CN aberration extent per chromosome (Table 2). To convert the feature space into counts that are suitable for decomposition, Macintyre *et al* utilised mixture models to categorise the feature distributions (Table 2). However, this mixture model approach may lead to different categories between different datasets or profiling platforms, causing difficulties when comparing results across different studies. To rectify this, Wang *et al* instead generated manual categories for each feature type, allowing for the same categories to be used across studies. A further study in multiple myeloma took a hybrid approach where categories for most features were determined using mixture models, but categories for CN were manually designed, specifically to fit the known biology of the malignancy [70].

Biological considerations of copy number signatures

Multiple studies have now demonstrated the utility of CN signatures to predict the prognosis of patients, both in a single-tumour [22,62,70] and in a pan-cancer

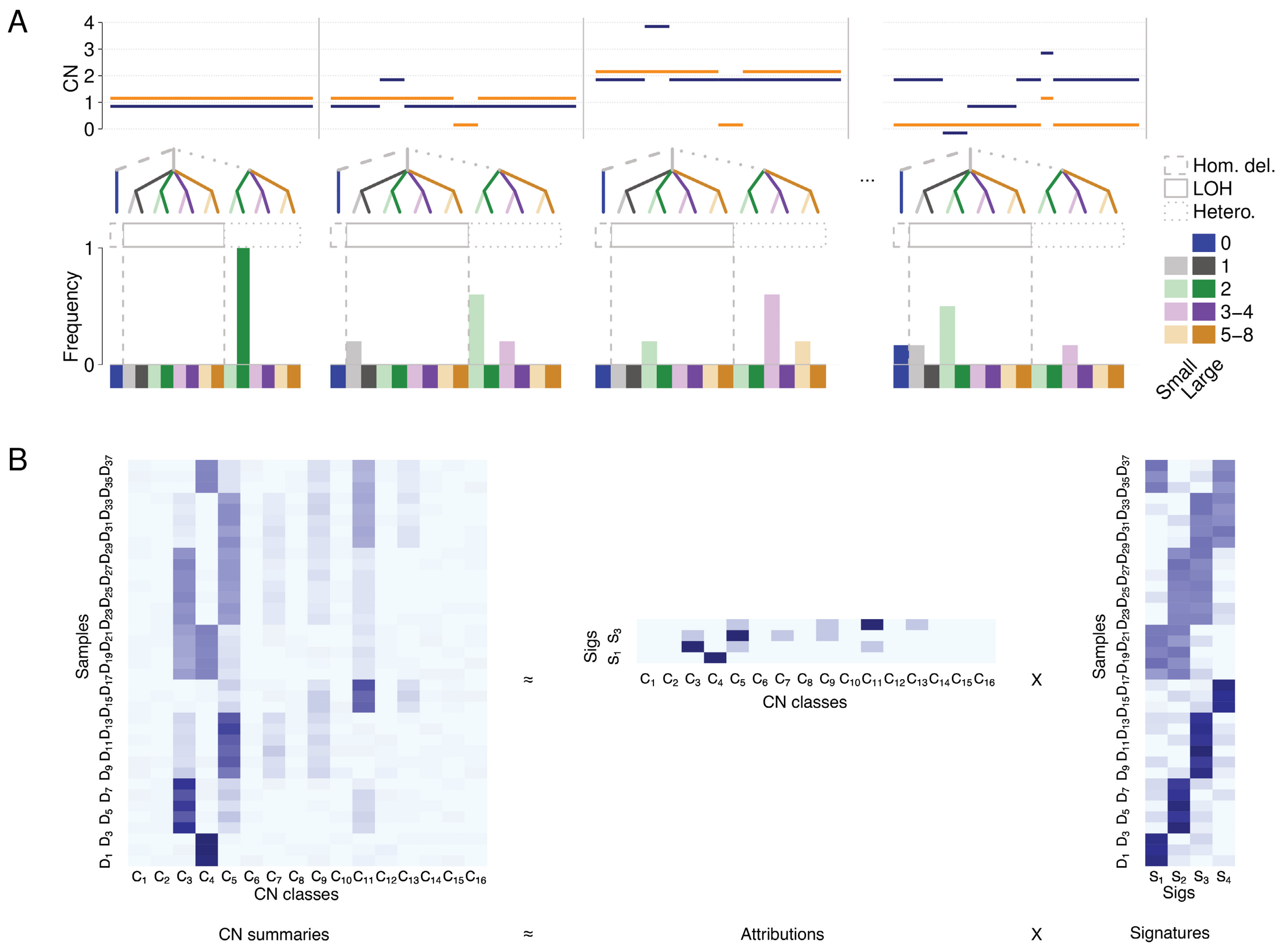

Figure 3. CN signature methodology (mechanism-agnostic). (A) Allele-specific CN profiles (top panels; *y*-axis = allele-specific CN, orange = minor CN, blue = major CN) have their segments classified by loss of heterozygosity (LOH) status, total CN, and segment size (middle panels; see legend in figure) to generate CN feature counts (bottom panels). Example profiles are shown for a diploid genome, a diploid genome with historic chromosomal instability, a genome double version of the previous genome, and a genome with extensive LOH (in order left to right). (B) CN feature counts for a dataset are combined into matrix of CN summaries (left), which are decomposed using non-negative matrix factorisation, or another appropriate method, into a matrix of signature definitions (right) and a matrix of attributions of each of those signatures in the samples of the given dataset (middle).

Table 2. Overview of published CN signature methods

	Macintyre *et al* (2018) [22]	Wang *et al* (2021) [62]	Maclachlan *et al* (2021) [70]	Steele *et al* (2019) [23]	Steele *et al* (2021) [24]
Tumour	Ovarian	Prostate	Multiple myeloma	USARC	Pan-cancer
Number of samples	385	1003	752	52	11 210
Number of signatures	7	5	5	7	21
Mechanism	Designed	Designed	Designed	Agnostic	Agnostic
Summary method	Mixture models	Manual categories	Hybrid (manual category – total CN)	Manual categories	Manual categories
Decomposition method	NMF	NMF	HDP	NMF	NMF
Number of components	36	80	28	40	48
Platforms	Shallow WGS	Exome	Shallow WGS	Deep WGS	SNP6 microarray Exome sequencing Deep WGS Shallow WGS Reduced representation bisulphite sequencing
Allele-specific	No	No	No	Yes	Yes
Genomic information	None	Chromosomal counts	None	None	*Post hoc* mapping
Included segments	All	All	All	All	All
Ploidy sensitive	Yes	Yes	Yes	Yes	Yes
Software	Custom code	Sigminer	Custom code	Custom code	SigProfiler

NMF, non-negative matrix factorisation; HDP, hierarchical Dirichlet process.

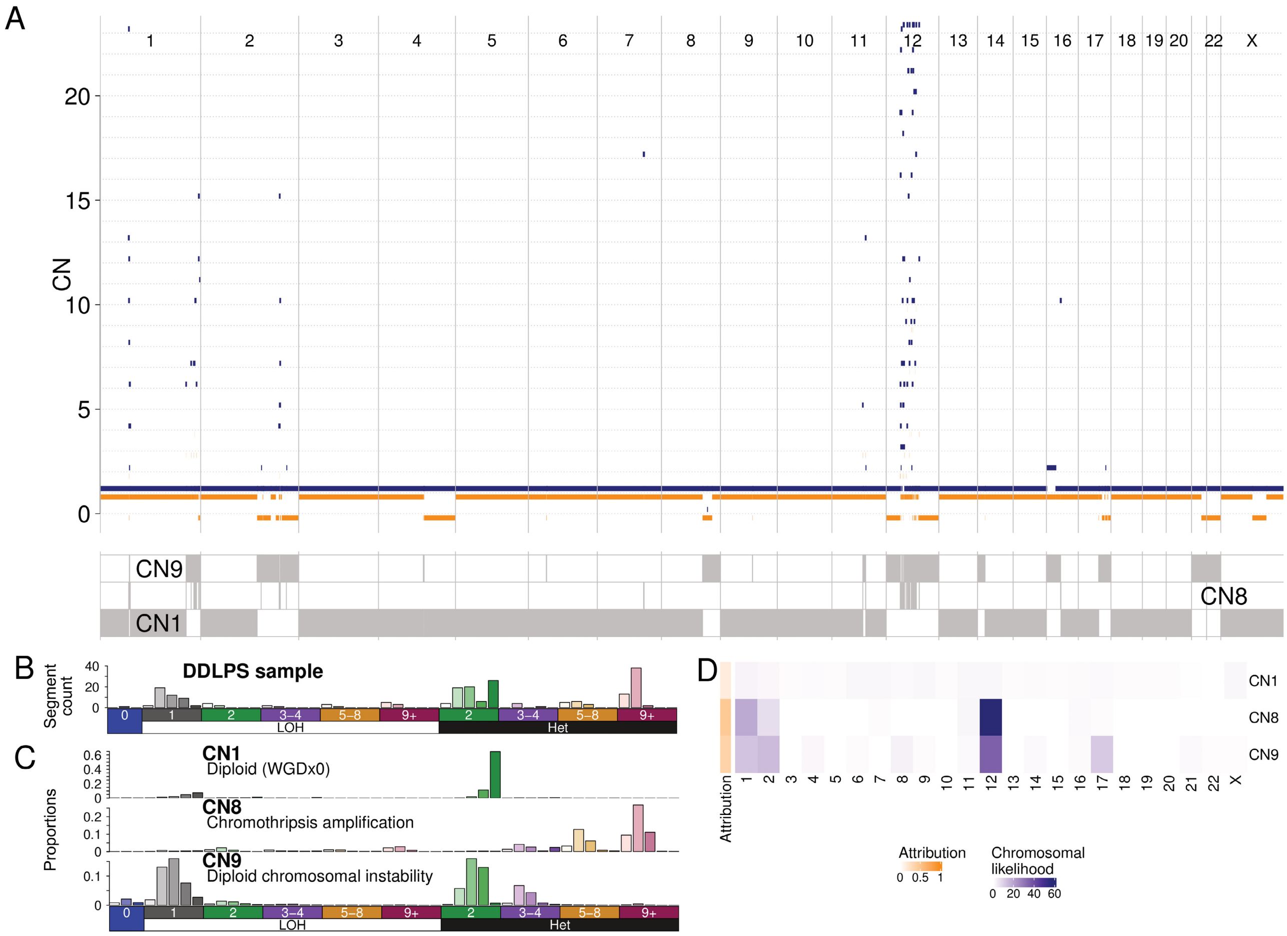

Figure 4. *Post hoc* mapping of CN signatures to the genome. (A) CN profile of a dedifferentiated liposarcoma (DDLPS) with characteristic amplification of chr12 including MDM2. Orange indicates minor CN; blue indicates major CN. The bottom panel indicates regions of the genome that have been mapped to a signature (in grey); regions of the genome unattributed to a signature are in white. (B) CN summary vector for the DDLPS in A. See legend of Figure 2 for the ordering of CN classes. (C) Signature definitions for the three signatures attributed to this sample. CN classes from B can be explained by each of the three signatures. (D) Heatmap of the likelihood of signatures (*y*-axis) explaining the segments in a genomic bin of the window (blue), where here the bins are the chromosomes of the human genome (*x*-axis). Global attributions of the three signatures to this sample are coloured in orange. Signature 8, a signature associated with chromothripsis amplification, is attributed to chromosomes 12, 1, and 2, due to the highly segmented patterns seen on those chromosomes. Once the chromosomal likelihoods are obtained, the maximum likelihood signature for each segment based on its CN class can be assigned, giving the assignments shown in the bottom panel of A.

context [24] (Table 2). Synthesis of the results to date suggests that patients with tumours that exhibit patterns indicating an amplicon or chromothripsis-like event have poor survival [24,70]. This corroborates previous findings of poor survival of patients with chromothriptic tumours [73] or with extra-chromosomal circular DNA [74].

The relative simplicity of the signature encoding in mechanism-agnostic approaches enables the mapping of signatures back to the genome, reconstituting the genomic context of the CN signatures [24], thereby providing information about the processes that generate the CN event. One notable example is the identification of chromothripsis signatures associated with oncogenic *MDM2* amplification in dedifferentiated liposarcoma (Figure 4) consistent with the recently described complex SV coupled with amplification known as *tyfonas* [30]. This result highlights the importance of genomic context when interpreting CN signatures, especially with the understanding that the CN landscape of a cancer cell is heavily shaped by positive and/or negative selection [75]. Beyond identifying known tumour suppressor genes/oncogenes using CN signatures, there may be further utility in discovering novel cancer driver genes, particularly in cohorts of rare or understudied tumours. The recurrence of specific signatures in regions of the genomes, including distinct recurrence patterns in individual tumour types [24], reflects the strong selective pressures acting on CNAs more broadly. This mirrors known phenomena such as recurrent aneuploidies seen in individual tumour types [76], driven through the interplay between loss of tumour suppressor genes, gain of oncogenes, and retention of essential genes [75]. In contrast, the vast majority of SBS mutations seen in a tumour genome are passenger events [77], and possibly have little effect on selection. This distinction suggests that SBS signatures provide a more unbiased window into the processes that generated them, whereas CN

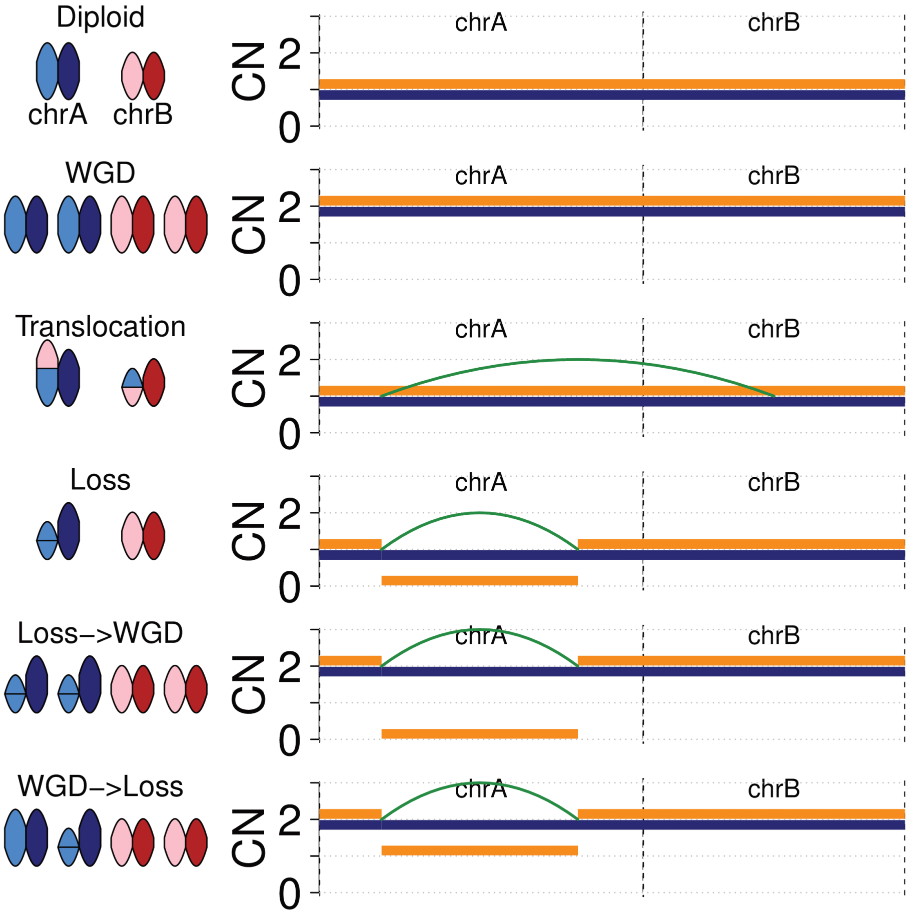

Figure 5. Schematic diagram of hypothetical genome configurations for two chromosomes (left; blue = chromosome A, red = chromosome B) and their associated CN and rearrangement profiles (*y*-axis = allele-specific CN, *x*-axis = genome, orange = minor CN, blue = major CN, green = rearrangement). Specific structural alterations to the genome may lead to only CN alterations (whole-genome doubling; WGD), only rearrangements (translocation), or a combination of both (loss). Further, the consequences of the same structural alterations (WGD and a loss) may depend on the order in which those structural alterations occurred; a loss followed by WGD (Loss → WGD) will lead to loss of heterozygosity for the lost segment, whereas the reverse ordering (WGD → Loss) will retain heterozygosity for that segment, which may be important when assessing second hits to tumour suppressor genes.

profiles have a much stronger inherent bias. This selection pressure has been directly observed in colon organoids, where daughter cells of chromosomal mis-segregation events have a high probability of subsequent cell death [78], whereas the same is not true of the vast majority of SBS mutations [27].

Future directions

CNAs and rearrangements represent two distinct but related consequences of structural alteration to the genome (Figure 5). As a result, it is likely that signatures that incorporate both rearrangements and CN profiles would more fully describe the consequences of structural phenomena. As an example, a whole-genome doubling event will be 'silent' when viewed through rearrangement but will be observed through CN data. In contrast, a balanced translocation or an inversion may be 'silent' within a CN profile but will be observed through rearrangement data. This may help to distinguish different molecular events, e.g. chromothripsis that has occurred on single versus on multiple chromosomes and may further refine patient stratification.

A distinction between CN changes and single base substitutions is that every substitution can be examined as a single independent event, barring local clusters of hypermutation [28] or violations of the infinite sites model [79,80]. In contrast, every segment in a CN profile cannot be evaluated as a single or as an independent event, as these CN events may occur simultaneously [81]. As a result, the interpretation of a CN profile is limited, in that you can only necessarily observe the end stage of a process that has generated the patterns. To exploit the potential of CN signatures for evolutionary studies, it may be valuable to perform sub-clonal CN reconstruction to discover the clonal composition of the tumour, and hence map CN events onto the tumour's evolutionary tree. In this way, it may be possible to develop a new generation of signatures of CNAs (event-level signatures) that may more accurately reflect the signatures of the observed CN profiles (end-state signatures). For SBS data, the difference between event-level and end-state-level is minimal [28], whereas for CN data it may be substantial.

Lastly, considerable effort has been expended to validate that signatures of small mutational events correspond to real biological processes [26,27,82,83]. Similar efforts will be required for CN signatures; however, the experimental strategies for inducing CNAs and/or processes are less deterministic and may necessitate careful experimental design. As a further complication, the CN profile generated from an individual event will depend on the previous history of the tumour genome, e.g. a loss before and after genome doubling will lead to loss or retention of heterozygosity, respectively (Figure 5). Despite these challenges, the emerging evidence suggests that CN signatures represent a fertile research area to understand cancer pathogenesis and can be utilised as a robust approach for prognostication and, in some cases, for therapeutic stratification of cancer patients.

Acknowledgements

NP holds a Cancer Research UK Clinician Scientist fellowship (Award No. 18387). CDS undertook this work with support from a Cancer Research UK Travel Award (Award No. 27969) and receives funding from the GEM Consortium (Boston Children's Hospital). Support was provided to NP by the University College London Hospitals Biomedical Research Centre and the Cancer Research UK University College London Experimental Cancer Medicine Centre. The Alexandrov laboratory is supported by US National Institutes of Health's R01 ES030993 and R01 ES032547. LBA is an Abeloff V Scholar and is supported by an Alfred P. Sloan Research Fellowship. Research at UC San Diego was also supported by a Packard Fellowship for Science and Engineering to LBA.

Author contributions statement

The review was conceived, designed, written, and edited by CDS, NP, and LBA.

References

1. Degtyareva NP, Saini N, Sterling JF, *et al*. Mutational signatures of redox stress in yeast single-strand DNA and of aging in human mitochondrial DNA share a common feature. *PLoS Biol* 2019; **17:** e3000263.
2. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* 2014; **24:** 52–60.
3. Schmeiser HH, Nortier JL, Singh R, *et al*. Exceptionally long-term persistence of DNA adducts formed by carcinogenic aristolochic acid I in renal tissue from patients with aristolochic acid nephropathy. *Int J Cancer* 2014; **135:** 502–507.
4. Parsons R, Li GM, Longley MJ, *et al*. Hypermutability and mismatch repair deficiency in RER$^+$ tumor cells. *Cell* 1993; **75:** 1227–1236.
5. Ganem NJ, Pellman D. Linking abnormal mitosis to the acquisition of DNA damage. *J Cell Biol* 2012; **199:** 871–881.
6. Griffith M, Miller CA, Griffith OL, *et al*. Optimizing cancer genome sequencing and analysis. *Cell Syst* 2015; **1:** 210–223.
7. Gerstung M, Jolly C, Leshchiner I, *et al*. The evolutionary history of 2,658 cancers. *Nature* 2020; **578:** 122–128.
8. Vogelstein B, Papadopoulos N, Velculescu VE, *et al*. Cancer genome landscapes. *Science* 2013; **339:** 1546–1558.
9. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458:** 719–724.
10. Morganella S, Alexandrov LB, Glodzik D, *et al*. The topography of mutational processes in breast cancer genomes. *Nat Commun* 2016; **7:** 11383.
11. Setlow RB, Carrier WL. Pyrimidine dimers in ultraviolet-irradiated DNA's. *J Mol Biol* 1966; **17:** 237–254.
12. Witkin EM. Ultraviolet-induced mutation and DNA repair. *Annu Rev Microbiol* 1969; **23:** 487–514.
13. Blum HF. On the mechanism of cancer induction by ultraviolet radiation. *J Natl Cancer Inst* 1950; **11:** 463–495.
14. Brash DE, Rudolph JA, Simon JA, *et al*. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc Natl Acad Sci U S A* 1991; **88:** 10124–10128.
15. Alexandrov LB, Kim J, Haradhvala NJ, *et al*. The repertoire of mutational signatures in human cancer. *Nature* 2020; **578:** 94–101.
16. Alexandrov LB, Nik-Zainal S, Wedge DC, *et al*. Signatures of mutational processes in human cancer. *Nature* 2013; **500:** 415–421.
17. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014; **15:** 585–598.
18. Abbasi A, Alexandrov LB. Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures. *DNA Repair (Amst)* 2021; **107:** 103200.
19. Koh G, Degasperi A, Zou X, *et al*. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* 2021; **21:** 619–637.
20. Li Y, Roberts ND, Wala JA, *et al*. Patterns of somatic structural variation in human cancer genomes. *Nature* 2020; **578:** 112–121.
21. Nik-Zainal S, Davies H, Staaf J, *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016; **534:** 47–54.
22. Macintyre G, Goranova TE, De Silva D, *et al*. Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* 2018; **50:** 1262–1270.
23. Steele CD, Tarabichi M, Oukrif D, *et al*. Undifferentiated sarcomas develop through distinct evolutionary pathways. *Cancer Cell* 2019; **35:** 441–456.e8.
24. Steele CD, Abbasi A, Islam SMA, *et al*. Signatures of copy number alterations in human cancer. *bioRxiv* 2021; 2021.2004.2030.441940 [Not peer reviewed].
25. Nik-Zainal S, Van Loo P, Wedge DC, *et al*. The life history of 21 breast cancers. *Cell* 2012; **149:** 994–1007.
26. Nik-Zainal S, Wedge DC, Alexandrov LB, *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* 2014; **46:** 487–491.
27. Petljak M, Alexandrov LB, Brammeld JS, *et al*. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* 2019; **176:** 1282–1294.e20.
28. Bergstrom EN, Luebeck J, Petljak M, *et al*. Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* 2022; **602:** 510–517.
29. Boot A, Liu M, Stantial N, *et al*. Recurrent mutations in topoisomerase IIα cause a previously undescribed mutator phenotype in human cancers. *Proc Natl Acad Sci U S A* 2022; **119:** e2114024119.
30. Hadi K, Yao X, Behr JM, *et al*. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* 2020; **183:** 197–210.e32.
31. Reijns MAM, Parry DA, Williams TC, *et al*. Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* 2022; **602:** 623–631.

32. Poon SL, Pang ST, McPherson JR, *et al*. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* 2013; **5:** 197ra101.
33. Hoang ML, Chen CH, Sidorenko VS, *et al*. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* 2013; **5:** 197ra102.
34. Dziubańska-Kusibab PJ, Berger H, Battistini F, *et al*. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat Med* 2020; **26:** 1063–1069.
35. Huang MN, Yu W, Teoh WW, *et al*. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res* 2017; **27:** 1475–1486.
36. Grolleman JE, de Voer RM, Elsayed FA, *et al*. Mutational signature analysis reveals NTHL1 deficiency to cause a multi-tumor phenotype. *Cancer Cell* 2019; **35:** 256–266.e5.
37. Chen H, Chong W, Teng C, *et al*. The immune response-related mutational signatures and driver genes in non-small-cell lung cancer. *Cancer Sci* 2019; **110:** 2348–2356.
38. Chopra N, Tovey H, Pearson A, *et al*. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. *Nat Commun* 2020; **11:** 2662.
39. Law EK, Sieuwerts AM, LaPara K, *et al*. The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Sci Adv* 2016; **2:** e1601737.
40. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 2020; **578:** 82–93.
41. Hutter C, Zenklusen JC. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* 2018; **173:** 283–285.
42. Martin CL, Warburton D. Detection of chromosomal aberrations in clinical practice: from karyotype to genome sequence. *Annu Rev Genomics Hum Genet* 2015; **16:** 309–326.
43. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* 1960; **25:** 85–109.
44. Arriola E, Marchio C, Tan DS, *et al*. Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines. *Lab Invest* 2008; **88:** 491–503.
45. Mitelman F, Johansson B, Mandahl N, *et al*. Clinical significance of cytogenetic findings in solid tumors. *Cancer Genet Cytogenet* 1997; **95:** 1–8.
46. Van Loo P, Nordgard SH, Lingjærde OC, *et al*. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 2010; **107:** 16910–16915.
47. Raine KM, Van Loo P, Wedge DC, *et al*. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr Protoc Bioinformatics* 2016; **56:** 15.9.1–15.9.17.
48. Kucharík M, Budiš J, Hýblová M, *et al*. Copy number variant detection with low-coverage whole-genome sequencing represents a viable alternative to the conventional array-CGH. *Diagnostics (Basel)* 2021; **11:** 708.
49. Boeva V, Popova T, Bleakley K, *et al*. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 2012; **28:** 423–425.
50. Carter SL, Cibulskis K, Helman E, *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012; **30:** 413–421.
51. Favero F, Joshi T, Marquard AM, *et al*. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015; **26:** 64–70.
52. Zaccaria S, Raphael BJ. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nat Commun* 2020; **11:** 4301.
53. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* 2021; **39:** 207–214.
54. Mermel CH, Schumacher SE, Hill B, *et al*. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011; **12:** R41.
55. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011; **474:** 609–615.
56. Wang Y, Ung MH, Cantor S, *et al*. Computational investigation of homologous recombination DNA repair deficiency in sporadic breast cancer. *Sci Rep* 2017; **7:** 15742.
57. Lord CJ, Ashworth A. BRCAness revisited. *Nat Rev Cancer* 2016; **16:** 110–120.
58. Xia F, Taghian DG, DeFrank JS, *et al*. Deficiency of human BRCA2 leads to impaired homologous recombination but maintains normal nonhomologous end joining. *Proc Natl Acad Sci U S A* 2001; **98:** 8644–8649.
59. Ledermann J, Harter P, Gourley C, *et al*. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N Engl J Med* 2012; **366:** 1382–1392.
60. Stover EH, Fuh K, Konstantinopoulos PA, *et al*. Clinical assays for assessment of homologous recombination DNA repair deficiency. *Gynecol Oncol* 2020; **159:** 887–898.
61. Glodzik D, Bosch A, Hartman J, *et al*. Comprehensive molecular comparison of *BRCA1* hypermethylated and *BRCA1* mutated triple negative breast cancers. *Nat Commun* 2020; **11:** 3747.
62. Wang S, Li H, Song M, *et al*. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet* 2021; **17:** e1009557.
63. Sztupinszki Z, Diossy M, Krzystanek M, *et al*. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer* 2018; **4:** 16.
64. Takaya H, Nakai H, Takamatsu S, *et al*. Homologous recombination deficiency status-based classification of high-grade serous ovarian carcinoma. *Sci Rep* 2020; **10:** 2757.
65. Davies H, Glodzik D, Morganella S, *et al*. HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat Med* 2017; **23:** 517–525.
66. Staaf J, Glodzik D, Bosch A, *et al*. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat Med* 2019; **25:** 1526–1533.
67. Nguyen L, Martens J WM, Van Hoeck A, *et al*. Pan-cancer landscape of homologous recombination deficiency. *Nat Commun* 2020; **11:** 5584.
68. Gulhan DC, Lee JJ, Melloni GEM, *et al*. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat Genet* 2019; **51:** 912–919.
69. Peng G, Chun-Jen Lin C, Mo W, *et al*. Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat Commun* 2014; **5:** 3361.
70. Maclachlan KH, Rustad EH, Derkach A, *et al*. Copy number signatures predict chromothripsis and clinical outcomes in newly diagnosed multiple myeloma. *Nat Commun* 2021; **12:** 5172.
71. Stephens PJ, Greenman CD, Fu B, *et al*. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011; **144:** 27–40.
72. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell* 2013; **152:** 1226–1236.
73. Cortés-Ciriano I, Lee JJ, Xi R, *et al*. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 2020; **52:** 331–341.
74. Kim H, Nguyen NP, Turner K, *et al*. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* 2020; **52:** 891–897.
75. Davoli T, Xu AW, Mengwasser KE, *et al*. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* 2013; **155:** 948–962.

76. Taylor AM, Shih J, Ha G, *et al*. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* 2018; **33:** 676–689.e3.
77. Van Hoeck A, Tjoonk NH, van Boxtel R, *et al*. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* 2019; **19:** 457.
78. Bolhaqueiro ACF, Ponsioen B, Bakker B, *et al*. Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids. *Nat Genet* 2019; **51:** 824–834.
79. Kuipers J, Jahn K, Raphael BJ, *et al*. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res* 2017; **27:** 1885–1894.
80. Demeulemeester J, Dentro SC, Gerstung M, *et al*. Biallelic mutations in cancer genomes reveal local mutational determinants. *Nat Genet* 2022; **54:** 128–133.
81. Umbreit NT, Zhang CZ, Lynch LD, *et al*. Mechanisms generating cancer genome complexity from a single cell division error. *Science* 2020; **368:** eaba0712.
82. Kucab JE, Zou X, Morganella S, *et al*. A compendium of mutational signatures of environmental agents. *Cell* 2019; **177:** 821–836.e16.
83. Zou X, Owusu M, Harris R, *et al*. Validating the concept of mutational signatures with isogenic cell models. *Nat Commun* 2018; **9:** 1744.