Genome Medicine

REVIEW

Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention

Song Ling Poon[1,2], John R McPherson[2,3], Patrick Tan[2,4,5,6], Bin Tean Teh[1,2] and Steven G Rozen[2,3*]

Abstract

Exposure to environmental mutagens is an important cause of human cancer, and measures to reduce mutagenic and carcinogenic exposures have been highly successful at controlling cancer. Until recently, it has been possible to connect the chemical characteristics of mutagens to actual mutations observed in human tumors only indirectly. Now, next-generation sequencing technology enables us to observe in detail the DNA-sequence-level effects of well-known mutagens, such as ultraviolet radiation and tobacco smoke, as well as endogenous mutagenic processes, such as those involving activated DNA cytidine deaminases (APOBECs). We can also observe the effects of less well-known but potent mutagens, including those recently found to be present in some herbal remedies. Crucially, we can now tease apart the superimposed effects of several mutational exposures and processes and determine which ones occurred during the development of individual tumors. Here, we review advances in detecting these mutation signatures and discuss the implications for surveillance and prevention of cancer. The number of sequenced tumors from diverse cancer types and multiple geographic regions is growing explosively, and the genomes of these tumors will bear the signatures of even more diverse mutagenic exposures. Thus, we envision development of wide-ranging compendia of mutation signatures from tumors and a concerted effort to experimentally elucidate the signatures of a large number of mutagens. This information will be used to link signatures observed in tumors to the exposures responsible for them, which will offer unprecedented opportunities for prevention.

New opportunities for detecting mutagen exposures in human tumors

Mutagenic environmental exposures are important causes of human cancer. This was first understood from Percival Pott's 18th century epidemiological observation of scrotal cancer in chimney sweeps [1]. Causality was eventually confirmed experimentally by using coal tar to induce cancer in rabbits [2]. Soon thereafter, polycyclic aromatic hydrocarbons were identified as carcinogens in coal tar [3]. Much later, once the role of DNA as an information molecule was understood, the biochemical mechanisms for polycyclic aromatic hydrocarbon mutagenesis were elucidated [4]. This led to a broader appreciation of the roles of DNA damaging agents in mutagenesis and to extensive study of numerous other mutagens [5,6]. Subsequently, assays for mutagenicity became proxies for tests of carcinogenicity, with the Ames test, performed in a bacterial system, as a well-known example [7]. However, tests of mutagenicity in artificial systems do not fully connect mutagenic exposures to the patterns of mutation observed in cancers.

More recently, it has become clear that specific mutagens produce characteristic patterns of somatic mutations in the DNA of malignant cells. We describe these patterns, called 'mutation signatures', in detail below. Briefly, mutation signatures usually include the relative frequencies of the various nucleotide mutations (such as A > C, A > G, A > T, C > A) plus, ideally, their trinucleotide contexts, that is, the identities of the bases on both sides of the mutated nucleotides. Previously, our knowledge of these signatures was based on short lengths (such as a few kilobases) of DNA sequence. With the advent of next-generation sequencing, it is now possible to infer these signatures from the sequences of all the exons in the genome ('whole exome') or from the sequence of the entire genome ('whole genome').

* Correspondence: steve.rozen@duke-nus.edu.sg
[2]Program in Cancer and Stem Cell Biology, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857, Singapore
[3]Duke-NUS Centre for Computational Biology, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857, Singapore
Full list of author information is available at the end of the article

Characterization of mutagenicity based directly on observed mutations across whole exomes or genomes offers several advantages over previous approaches, including that many more mutations can be detected, which provides far greater statistical power and allows the parsing of the superimposed mutation signatures stemming from several exposures. Actual mutation signatures are the end result of a series of biochemical and biological processes, including the metabolism of pro-mutagens to active forms, biochemical damage to DNA, the efforts of the cell to repair the damage, and, rarely, selection for or against the resulting mutations. Thus, while not obviating the need for mechanistic studies of the biochemical mechanisms of mutagenicity, cataloging mutations by next-generation sequencing provides information about a critical endpoint: the actual mutations that occur in cell lines or in human cancers in response to mutagenic exposures.

The long-term promise is that the epidemiological connection of specific mutagens to signatures actually observed in tumors will indicate which mutagenic exposures are true substantial contributors to the burden of human cancer (Figure 1).

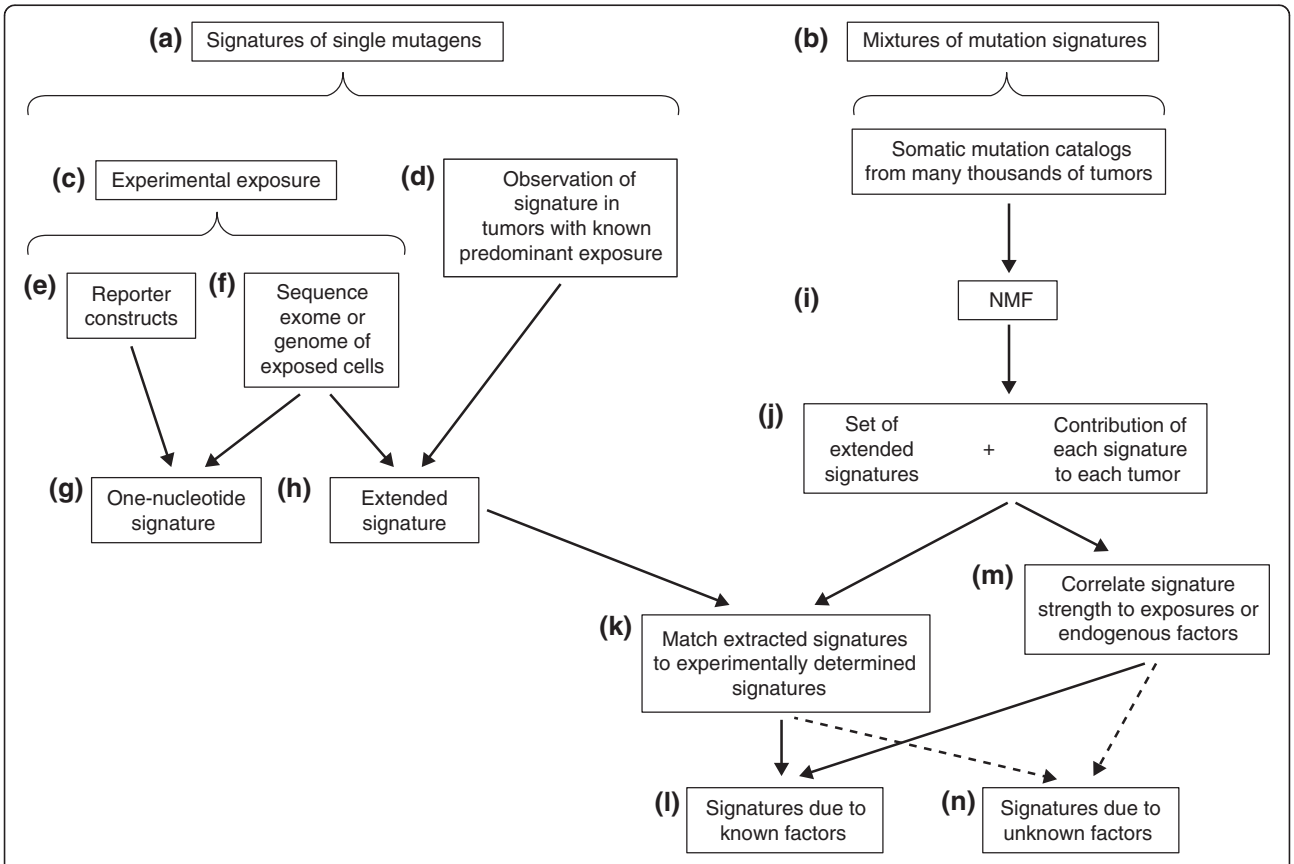

Figure 1 Linking mutation signatures to exposures or endogenous mutational processes. One can either **(a)** focus on signatures of one mutagen at a time or **(b)** study mixtures of signatures. One can study signatures of single mutagens either **(c)** via experimental approaches, or **(d)** via observation of mutation signatures in the exome or genome sequence of tumors with a known predominant mutational exposure. Some tumor exomes harbor only a handful of somatic point mutations, and presumably these tumors arise from causes other than mutagenesis. For many cancers, typical numbers of somatic point mutations in exomes are 60 to 300 [9-12]. Highly mutated cancers sometimes have >3,500 mutations per exome [13]. Typical numbers for genomes of cancers such as those of the lung or stomach are >15,000 [14,15], and a few highly mutated genomes harbor >400,000 somatic point mutations [16]. Among experimental approaches, one can use **(e)** reporter constructs and observe mutations in short sequences. This allows inference of relatively simple signatures, for example **(g)** signatures involving only single nucleotide mutations. **(f)** By sequencing the exome, or, ideally, the genome of a mutagen-exposed, clonal cell line, one can **(h)** infer a more informative, extended signature, for example one that includes the trinucleotide contexts of single-nucleotide mutations. One can also infer extended signatures by sequencing the exomes or genomes of tumors with known, predominant exposures **(d)**. Extraction of mutation signatures from mixtures of signatures **(b)** requires somatic mutation catalogs from the exomes or genomes of large numbers of tumors. The most recent studies have looked at thousands of catalogs. **(i)** Procedures based on NMF (non-negative matrix factorization) allow **(j)** simultaneous inference of a set of extended mutation signatures and the contributions of each inferred signature to each tumor's mutations. **(k)** Extended signatures derived from mixtures of signatures **(j)** can be matched to extended signatures that were experimentally determined or inferred from known predominant exposures **(h)**, thereby providing information on **(l)** exposures that contributed to tumors with mixtures of mutation signatures. **(m)** Alternatively, extended signatures extracted from mixtures can be correlated with information on mutagenic exposures or on endogenous mutagenic factors, allowing inference of the causes of mutation signatures **(l)**. **(n)** The causes of some signatures will remain unknown and require further research.

The observation of known signatures in tumors might also implicate previously unsuspected exposures in particular cancers. This information on causal exposures then could provide foci for prevention efforts. Nevertheless, despite the low cost and ubiquity of next-generation sequencing, detailed mutation signatures of only a few known carcinogens have been elucidated experimentally so far. Indeed, in a recent groundbreaking survey of mutation signatures across many types of human cancers, most mutation signatures are ascribed to particular exposures by statistical association rather than recapitulation of the signatures in experimental systems [8].

We describe below the state of the art for determining mutation signatures by next-generation sequencing, the implications of this approach for detecting the carcinogenic impacts of mutagenic exposures, and its promise for prevention. We start by describing signatures of single mutagens. We then describe approaches for teasing apart superimposed signatures from multiple mutagenic processes, and conclude with a vision of how this could improve prevention.

Signatures of single mutagens

To date, the signatures of carcinogenic mutagens have been established either *in vitro* or in human cancers that are primarily caused by one exposure (Table 1). We elaborate first on the mutation signature of aristolochic acid (AA), which has been established both *in vitro* and in human cancers [16-19]. AA is a powerful mutagen that is found in some herbal remedies and that causes upper urinary-tract urothelial cancer (UTUC) [16-19]. It also probably contributes to liver cancer (hepatocellular carcinoma, HCC) [16]. Thus, in addition to providing an example of the signature of a single mutagen, AA also illustrates the use of signatures to detect likely carcinogenic exposures that were previously unsuspected.

Table 1 Examples of exogenous mutagens and endogenous mutagenic processes

Mutagen	Dominant mutations*	Extended context*	Studies reporting mutation signatures	Prevalence	Challenges
Exogenous					
Aristolochic acid	A > T	(C|T)AG > (C|T)TG	[16-19]	Widely used in traditional medicines [20]; exposure to AA is widespread in Taiwan [17]	No unusual challenges
UV radiation	C > T;strand bias; CC > TT	TC > TT (C|T)C > (C|T)T	[6,8,9,21-23]	Prevalence of signature: 87% of melanoma [8]	No unusual challenges
Tobacco smoke	Primarily C > A, some C > G and C > T	CG > AGCG > TG;CG > GG	[8,14,24-26]	Extensive epidemiological evidence of the role of tobacco smoke in cancer [27,28]	Contains multiple carcinogens with individually unknown signatures
Aflatoxin B1	Primary G > T; some G > A	NA	[29-34]	[29,35]	Signature in extended context not known
Temozolomide	C > T	CC > TC; CT > TT	[8,36,37]	Present in 10% of glioblastomas; 9% of melanoma [8]	No unusual challenges
Benzene	C > T;C > A	NA	[38]	Exposure associated with risk of leukemia [39,40]	Several mutagenic metabolites and signature in extended context not known
Endogenous					
Activated APOBEC	C > T	TCA > TTA	[8,41,42]	Present in 16 tumor types [8]	Signatures 2 and 13 are similar [8], except 2 has C > T and 13 has higher C > G
Mutated DNA polymerase epsilon	C > T	TCG > TTG; TCT > TAT	[8,13]	Present in 13.7% of uterus cancer and 36.7% of colorectal cancer [8]	No unusual challenges
Mismatch repair deficiency (MSI)	C > T;C > A	CG > TG; CT > AT; homopolymer and microsatellite length changes	[8]	Present in 9 tumor types [8]	No unusual challenges
Correlated with patient age	C > T	CG > TG	[8]	A majority of tumors of most types have this signature [8]	Interpretation of two similar signatures in [8] not clear

*The ' > ' symbol indicates a change from one nucleotide to another; a vertical line indicates alternative nucleotides. NA, not applicable.

Mutation signature of aristolochic acid

AA is a natural compound found in plants in the genus *Aristolochia* (Figure 2a). These plants are used in traditional herbal remedies for weight loss and a plethora of health problems, including menstrual symptoms, snakebites, rheumatism, arthritis, and gout [43,44]. Although challenging to document, the use of these plants probably remains widespread [20,45]. Indeed, 99 species in the genus are known to be used medicinally, and although the AA content of most species is unknown, 23 of the medicinally used species contain AA [46]. AA is metabolized to aristolactam nitrenium ions, which form covalent adducts with adenosines in DNA (Figure 2b-e) [45,47]. These adducts then lead to A > T mutations (mutations from adenine to thymine; Figure 2f). These were initially observed as somatic mutations in the

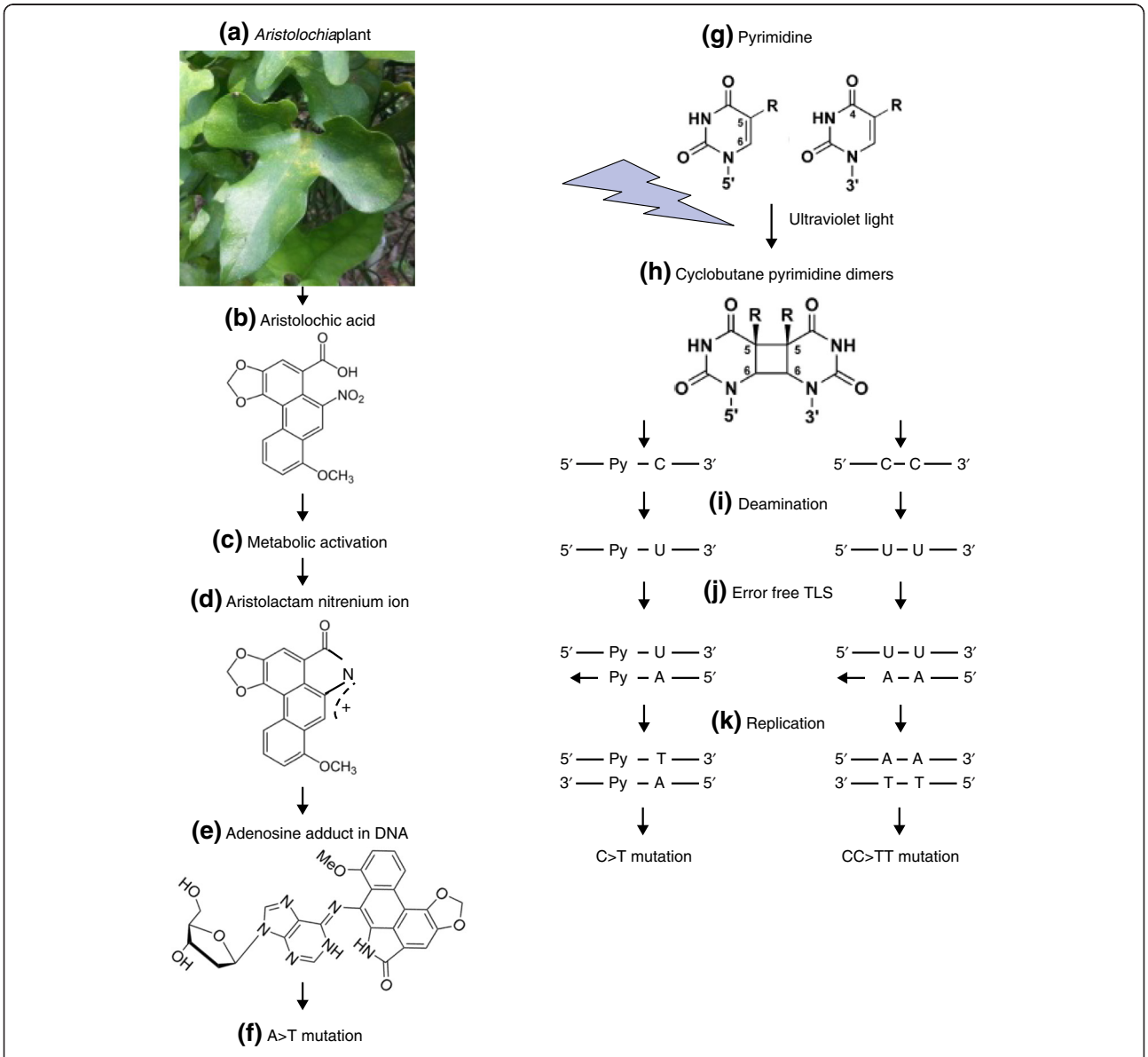

Figure 2 Mechanisms of mutagenesis of aristolochic acid and UV light. Preparations of **(a)** plants from the genus *Aristolochia* contain **(b)** aristolochic acid. Aristolochic acid I is shown; in aristolochic acid II, OCH_3 is replaced by H. Aristolochic acid is **(c)** metabolically activated to **(d)** aristolactam nitrenium ions by one or more of several enzymes, including NQO1 (NAD(P)H dehydrogenase, quinone 1), CYP1A2 (cytochrome P450, family 1, subfamily A, polypeptide 2), and NADPH-hemoprotein reductase [45]. **(e)** The aristolactam nitrenium ions form covalent adducts with adenosine bases, and **(f)** these adducts lead to A > T mutations. **(g)** Pyrimidines exposed to ultraviolet (UV) radiation form **(h)** cyclobutane pyrimidine dimers (CPD). **(i)** Either the cytosine (C) (left) or the CC dipyrimidines in CPD (right) undergo deamination, resulting in uracil (U). Py denotes pyrimidine. **(j)** Error-free trans-lesion DNA synthesis (TLS) induces C > T and CC > TT mutations at the sites of U-containing CPDs through DNA replication of the U-containing DNA strand **(k)**. Photograph of *Aristolochia* plant **(a)** by ST Pang. Molecule schematics in **(b-e)** reproduced from [14], with permission from Oxford University Press.

tumor suppressor gene *TP53* in UTUCs in the UK, Taiwan, and the Balkans [17,47]. This was highly unusual, as A > T mutations are rare in other types of human cancer, including UTUCs unrelated to AA exposure [18]. Furthermore, the AA-associated mutations in *TP53* tended to occur in the context of CAG > CTG (C followed by the mutated A followed by G, in 5' to 3' order) [19].

However, analysis of approximately 1 kb of sequence in a single gene (*TP53*) [19] offers limited statistical power to determine the sequence contexts in which the A > T mutations occur. In addition, the approach of assessing physical mutation signatures in *TP53*, a key tumor suppressor gene, runs the risk of bias caused by conflation of physical mutation signatures with the effects of intense selection during tumor evolution.

Recently, high-throughput next-generation sequencing has provided the means to catalog and analyze somatic mutations far more completely, whether by whole-exome or whole-genome sequencing. Recent work has shown a remarkable preponderance of A > T mutations in AA-associated UTUCs from Taiwan (Figure 3a) [16-19]. For comparison, in gastric cancer or other non-AA-associated cancers, A > T somatic mutations are rare (Figure 3b) [8,14,24].

By way of technical explanation, if we consider a single DNA strand as a point of reference, there are 12 possible single-nucleotide mutations: four nucleotides times three possible mutations for each nucleotide. In some parts of the genome, it makes sense to use a particular strand as the reference sequence. In particular, in regions of the genome that are transcribed, we can use the transcribed strand, that is, the strand that serves as a template for the RNA polymerase, as the point of reference. However, in the non-transcribed regions, neither strand in particular is the obvious choice for the reference sequence. Therefore, the usual practice in the study of mutation signatures has been to not distinguish complementary mutations, but rather to group them together. For example, A > C mutations are grouped with the complementary T > G mutations, A > G mutations are grouped with T > C mutations, and so on.

With the availability of catalogs of somatic mutations from sequencing data, it has become possible to investigate the nucleotides that neighbor AA-induced A > T mutations. The trinucleotide sequence contexts of AA-associated mutations show a dramatic overrepresentation of cytosines and thymines immediately 5′ of mutated adenines (that is, [C|T]**A**; mutated adenine in bold) and overrepresentation of guanines 3′ of mutated adenines (that is, **A**G) (Figure 3c) [16-19]. This preference of A > T mutations for the (C|T)**A**G context has not been observed in non-AA-associated cancers (such as gastric cancer; Figure 3d), suggesting that this sequence context is a particular characteristic of AA mutagenesis.

In addition, the A > T mutations in AA-associated UTUCs are less common on the transcribed strands of genes than on the non-transcribed strands (Figure 3e). This strand bias suggests that AA adducts occurring on the transcribed strand were often corrected by transcription-coupled nucleotide excision repair. Similar strand bias is not seen for the relatively infrequent A > T mutations seen in other cancers, such as gastric cancer (Figure 3f).

An AA-like signature in liver cancer

Unexpectedly, recent examination of mutation signatures in hepatitis B virus-exposed human HCCs revealed some with obvious AA-like signatures (Figure 3g,h) [16], although this cancer type apparently was not previously linked to AA exposure [48]. The signature shows a large proportion of A > T somatic mutations with strand bias (as seen in AA-exposed UTUCs; Figure 3e,g) and a trinucleotide context that strongly resembles that in AA-associated UTUC (compare Figure 3c and Figure 3h). It is possible that exposure to AA in conjunction with hepatitis B virus infection may contribute synergistically to HCC formation, much as hepatitis and aflatoxin do (see below). As AA had not been previously implicated as a risk factor for HCC, this finding may represent a new paradigm, in which environmental exposures contributing to specific cancers are deduced from observations of mutation signatures. It is likely that *Aristolochia*-containing herbal remedies are the source of AA exposure in these cancers. If so, appropriate measures to minimize exposure should be taken - for example, through education and more aggressive enforcement of bans on *Aristolochia*-containing remedies.

Ultraviolet radiation

Ultraviolet (UV) radiation induces several kinds of mutations, primarily C > T (Figure 2g-k, Table 1) [6,9]. It also induces double mutations CC > TT, in which adjacent cytosines mutate to thymines as a result of cytosine dimers generated by UV light. Earlier studies indicated that UV-induced C > T mutations often occur after a pyrimidine (C or T) [9,21,22]. Analysis of mutation catalogs from melanomas indicates that the trinucleotide context is often TCC [8]. As with AA-induced A > T mutations, there is strand bias: UV-induced mutations are less likely to occur on the transcribed strand [8].

Tobacco smoke

Tobacco smoking causes the vast majority of lung cancers and contributes strongly to many other cancers, including liver, colorectal, breast, prostate, and bladder cancers [49]. Tobacco smoke contains many mutagenic carcinogens, including polycyclic aromatic hydrocarbons and N-nitrosamines [25,50,51]. The mutation signature

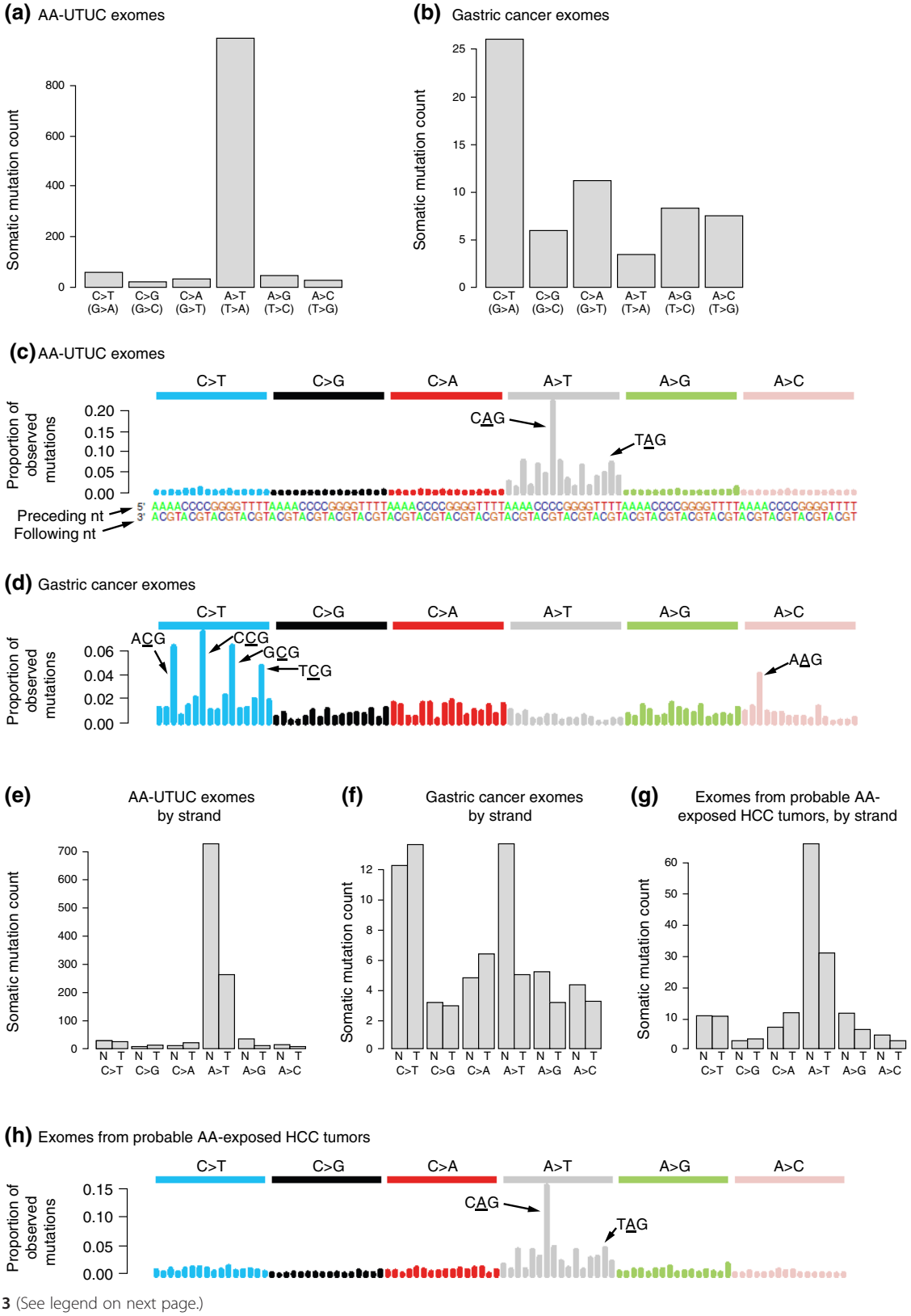

Figure 3 (See legend on next page.)

(See figure on previous page.)
Figure 3 Aristolochic acid signatures in upper urinary-tract urothelial cancer and hepatocellular carcinoma. (a,b) Mean counts of each of six different somatic single-nucleotide mutations in exome data from **(a)** AA-associated UTUCs (AA-UTUC, $n = 9$) and **(b)** gastric cancers ($n = 15$). **(c,d)** Trinucleotide contexts for somatic mutations in **(c)** AA-UTUCs ($n = 9$) and **(d)** gastric cancers ($n = 15$). The height of each bar (the *y* axis) represents the proportion of all observed mutations that fall into a particular trinucleotide mutational class, for example CAG > CTG and TAG > CTG (indicated). Along the *x* axis the mutations are organized first by the nucleotide mutation itself: C > T (blue bars), C > G (black bars), C > A (red bars), A > T (gray bars), A > G (green bars), A > C (pink bars). For each single-nucleotide mutation (such as A > T) there are 16 possible trinucleotide contexts (AAA > ATA, AAC > ATC, and so on) The heights of the bars indicate the observed proportions of mutations aggregated over all exomes studied. **(e,f)** Mean counts of somatic single-nucleotide mutations in **(e)** AA-associated UTUCs ($n = 9$) and **(f)** gastric cancer ($n = 15$), shown separately for non-transcribed (N) and transcribed (T) strands. The lower mutation counts on the transcribed strand suggest transcription-coupled repair (see main text). **(f)** Analogous data for gastric cancer do not show strand bias ($n = 15$). **(g)** Probable AA-exposed HCCs show a preponderance of A > T mutations with strand bias similar to that observed in AA-associated UTUCs ($n = 11$). **(h)** Trinucleotide context for mutations in probable AA-exposed HCC is highly similar to that for AA-associated UTUCs **(c)**. Plotted using data from [16].

of tobacco smoke was studied primarily in the context of the *TP53* gene, in which exposure to tobacco-smoke mutagens often results in G > T mutations [25]. Only a few studies extended the mutation signature to a trinucleotide context, and the preference for particular nucleotides 5' or 3' of the mutated nucleotides is weak (Table 1) [8,24], possibly reflecting the complex mix of mutagens present in tobacco smoke. There are challenges in dissecting the tobacco-smoke mutation signature, because the signatures from different constituent mutagens are likely to differ, and their effects on different organs and tissues are also likely to differ [51]. Thus, it would be highly informative to examine experimentally the signatures of individual mutagenic components of tobacco smoke in the genomes of exposed cell lines from different tissues (Figure 1f).

Aflatoxin B1

Aflatoxins are byproducts of mold growing on food [52], and among the aflatoxins, aflatoxin B1 (AFB1) is thought to be the most carcinogenic and is the most studied [53]. The International Agency for Research on Cancer (IARC) classifies AFB1 as a Group I carcinogen (an agent that is definitely carcinogenic to humans) [54]. AFB1 is metabolized to an epoxide compound that can form a covalent bond with the N7 atom of guanine, thereby leading to G > T mutations (Table 1) [35]. In addition, AFB1 can induce 8-hydroxy-2'-deoxyguanosine, which also produces predominantly G > T mutations in *in vitro* experimental models [52]. The mutation signature of AFB1 has been primarily studied in the *TP53* gene, and indeed particular somatic mutations in *TP53* are used as biomarkers for aflatoxin exposure in tumors [55,56]. However, the extended mutation signature of AFB1 has not been studied (Table 1). Exposure to AFB1 is through food, but unfortunately, its contamination in food is difficult to detect. Consequently, convincing evidence that AFB1 is carcinogenic relied on studies showing that people with AFB1-derived adducts were more likely to develop cancer [29,35]. The predominant cancer associated with AFB1 is HCC, and the risk associated with combined AFB1 exposure and hepatitis infection is far greater than each individual risk [29,35].

Temozolomide

Temozolomide is an alkylating agent commonly used for chemotherapeutic treatment of melanoma and central nervous system tumors [57,58]. Temozolomide is quickly absorbed and undergoes spontaneous breakdown to form an active compound (methyltriazen-1-yl imidazole-4-carboxamide), which forms several DNA adducts: *N*7-methylguanine (70%), *N*3-methyladenine (9%), and *O*6-methylguanine (5%) [59]. Both the *N*7-methylguanine and *N*3-methyladenine lesions are rapidly repaired by base excision repair [60]. However, the *O*6-methylguanine adducts sometimes are not repaired, leading to point mutations [61,62]. Although the mechanisms of temozolomide genotoxicity have been intensively studied in a therapeutic context, to our knowledge, the mutation signature of temozolomide has not been studied in experimental systems. However, Alexandrov *et al.* [8] detected a clear association between a CC > TC signature and temozolomide treatment in glioblastoma and melanoma patients (Table 1).

Benzene

Occupational exposure to benzene is of particular concern, as it is widely used in a variety of industries, including manufacture of petrochemicals and other chemicals, as well as in manufacture of shoes, lubricants, dyes, detergents, drugs, and pesticides [63]. Non-occupational exposures occur from automobile exhaust and gasoline fumes, industrial emissions, and especially cigarette smoking and second hand smoke [63]. Benzene is classified as a Group 1 carcinogen by IARC [64]. It is benzene's metabolites, such as phenol, hydroquinone, and related hydroxyl metabolites, that have been linked to leukemia in experimental models *in vitro* and *in vivo* [65,66]. Benzene metabolites can exert their genotoxic effect through the formation of DNA adducts, oxidative stress, damage to the mitotic apparatus, and inhibition of topoisomerase II function [65]. Although the

genotoxic mechanisms of benzene have been studied, its mutation signature is poorly understood. Thus far, research using a reporter gene has found a preponderance of C > T and C > A mutations [38] (Table 1). However, there has been no genome-wide analysis of benzene's mutation signature in cell line models or in benzene-associated leukemias.

Endogenous mutagenic processes

There are also endogenous mutagenic processes, which are sometimes unleashed during cancer development. For example, the *APOBEC* genes encode DNA cytidine deaminases that, when upregulated, promote C > G and C > T mutations especially in the T**C**(A|T) context (mutated base in bold; Figure 4a, Table 1) [41,42,67]. Endogenous mutagenic processes arising in cancer development can also consist of inactivation of DNA repair or proofreading mechanisms. A well-known example is microsatellite instability, caused by defects in the DNA mismatch repair mechanism [68]. As another example, it was recently shown that, in some cancers, inactivation of the proofreading domain of DNA polymerase delta 1 or epsilon (*POLD1* or *POLE*) leads to very high mutation rates [13]. *POLE* mutations were associated with very high rates of TCT > TAT and TCG > TTG mutations [8] (Table 1).

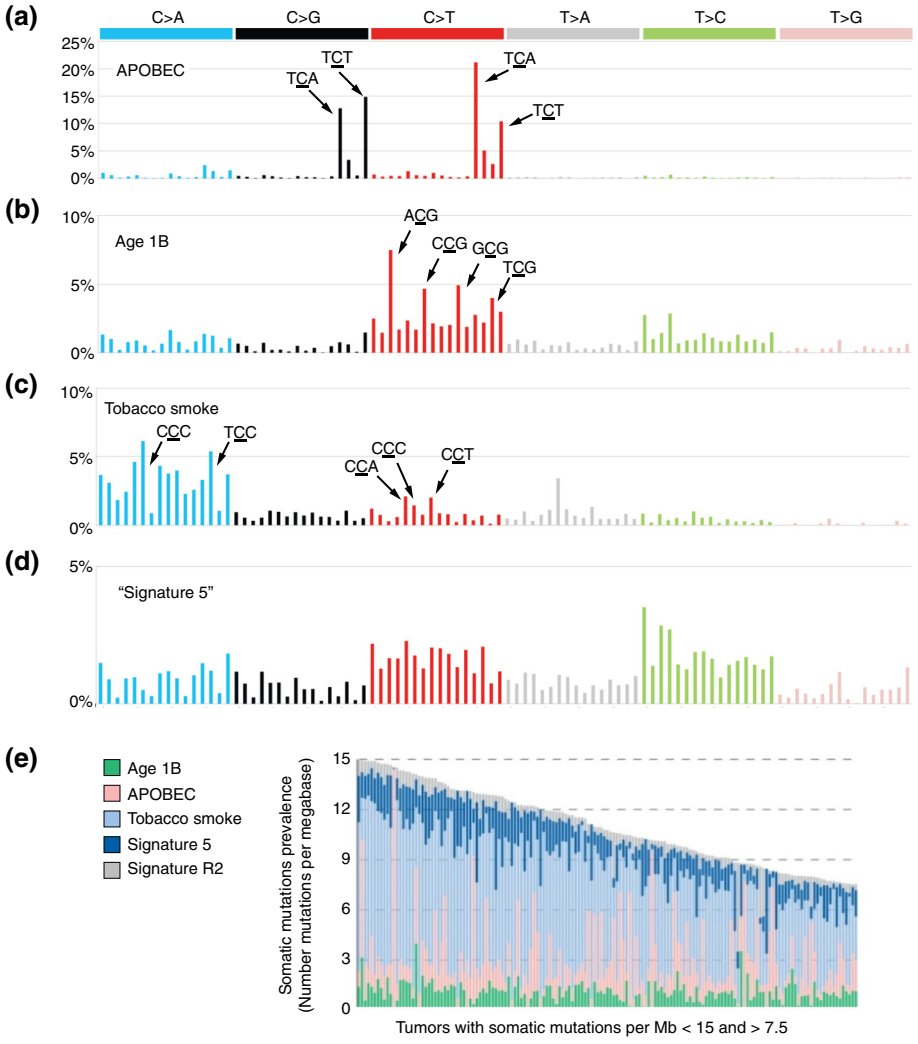

Figure 4 Mutation signatures in lung adenocarcinoma. (a-d) Four signatures that are prominent in lung adenocarcinoma [8]. The *x* axes are organized as in Figure **3**c. However, unlike in Figure **3**c, the *y* axes in these plots represent proportions of mutations in inferred rather than observed signatures. **(a)** Signature of APOBEC-induced mutagenesis. **(b)** 'Age 1B', one of two signatures that correlated with age. **(c)** Tobacco smoke. **(d)** 'Signature 5' from [8], due to an unknown exposure or mutational process. **(e)** Almost all lung adenocarcinomas have mutations that are overlays of several of the signatures above. Signature R2 is an additional signature that may partly represent sequencing errors [8]. All panels adapted from [8] with permission from Macmillan Publishers Ltd.

Aging by itself is a major risk factor for cancer development, and the majority of tumors are diagnosed in older patients [69-71]. DNA damage and mutations accumulate with age [72]. Interestingly, there are different age-related mutation patterns in different tissues due to differences in functional characteristics such as mitotic rate, transcriptional activity, metabolism, and specific DNA repair mechanisms [73]. Two distinct yet similar age-related mutation signatures have been detected in cancers (Table 1), and at least one of the two is present in the overwhelming majority of tumors [8].

Mixtures of signatures

In most tumors, somatic mutation catalogs comprise the superimposed results of several mutational exposures and processes. For example, lung adenocarcinomas usually show the signature of tobacco smoke [8,14,24] (Figure 4c). In addition, these tumors often simultaneously show mutation signatures due to exposure to endogenous activated DNA cytidine deaminases (APOBECs; Figure 4a), signatures of mutations that accumulate with age (Figure 4b), and other signatures of unknown origin [8] (Figure 4d).

Given that the catalog of somatic mutations in a tumor often represents an overlay of several mutational processes, a key challenge is to dissect out and assess the contribution of each process. Building on initial work [74], recent strides have been made in computational techniques for meeting this challenge. Specifically, it is now possible to simultaneously discover the existence of multiple signatures and assess the relative contribution of each signature to each tumor's catalog of somatic mutations [8,75,76]. Figure 5 explains the process of combining three mutation signatures to reconstruct a close approximation to the observed somatic mutation catalog of a tumor.

Discovering the signatures relies on a computational analysis called non-negative matrix factorization (NMF). The input to NMF consists of the observed catalogs of somatic mutations from tens [75] to several thousands [8] of tumors. For each of the observed catalogs (one for each tumor), NMF sets up an equation such as the one shown in Figure 5. Then, for a pre-specified number, N, of undefined component signatures, NMF finds the N specific signatures and the contributions of each specific signature (the 'pie chart' circle, Figure 5b) that, for all the tumors simultaneously, provide the closest reconstructions of the observed catalogs. In its mathematical formulation, the collection of mutation catalogs (Figure 5d) is the approximate product of the matrix representing the mutation signatures (Figure 5a) and the matrix representing the contributions of each signature to each tumor (Figure 5b). In other words, Figure 5a and Figure 5b are factors that, when multiplied, yield an approximation of Figure 5d. These factors are constrained to be non-negative, because one cannot have a negative contribution of a mutation signature to a tumor, and because a mutation signature cannot have a negative proportion of mutations of a given class; this is the origin of the term non-negative matrix factorization. We emphasize that NMF simultaneously detects the signatures present in the somatic mutation catalogs of multiple tumors and determines the contribution of each signature to the somatic mutations in each tumor.

There are, of course, numerous fine points, salient among which is the question of how to find the right number, N, of signatures. This depends on the number of mutation catalogs (and the number of mutations) available for analysis, as well as on the actual diversity of mutational processes represented in the sampled tumors. A large international effort recently generated somatic mutation catalogs from 7,042 tumors encompassing 30 cancer types, and these catalogs allowed discernment of 21 mutation signatures [8]. Across all the tumors analyzed, every cancer type had at least two mutation signatures; the cancers with the most signatures were those of the liver (seven signatures) and stomach and uterus (six signatures each). Figure 4e shows the example of lung adenocarcinomas, which usually show mixtures of several mutational processes.

Based on association with clinically documented exposures or correspondence to previously known mutational profiles, the origins of 11 of the 21 signatures in [8] were identifiable. Three signatures were attributed to exogenous exposures: tobacco smoke, UV radiation, and temozolomide. Other signatures were attributed to endogenous processes, including activation of *APOBEC* genes, mismatch repair deficiency, mutations in the *POLE* gene, and mutations in the *BRCA1* or *BRCA2* breast cancer genes. Finally, there were two signatures for which the level of the contribution to mutations in tumors was strongly correlated with the patient's age.

Despite the power conferred by analysis of mutation signatures across the 7,042 tumors, the environmental or biological factors underlying 10 of the 21 signatures could not be identified, and indeed only three signatures were linked to exogenous exposures [8]. Furthermore, over two-thirds of the cancer types studied harbored signatures of unknown source. Thus, there is a large gap in our understanding of the environmental exposures and mutational processes that contribute to common human cancers. Conversely, there are mutagens with well-studied biochemistry - for example, aflatoxins [30], benzene [66], and AA - that were not detected in these tumors. Possibly none or few of the 7,042 tumors analyzed had been exposed to these mutagens. Indeed, it seems likely that none were exposed to AA, which has a very distinctive signature that would have been detected

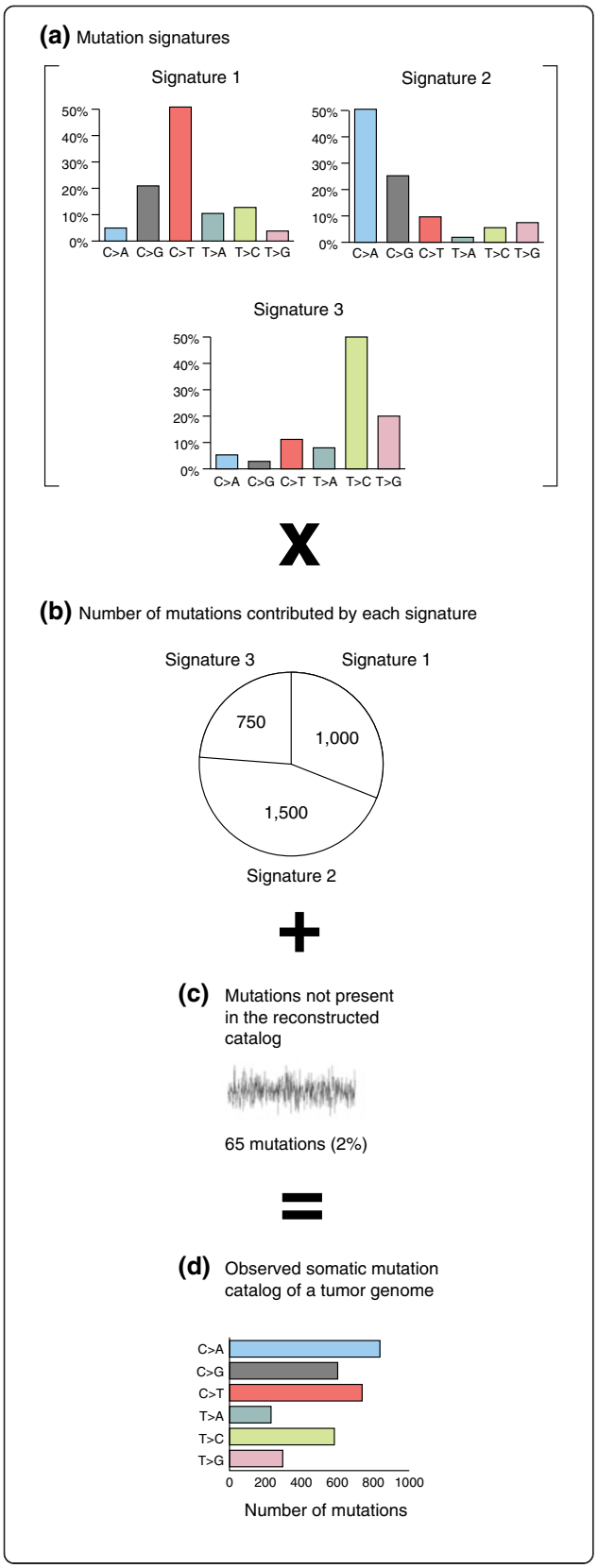

Figure 5 Reconstructing the catalog of somatic mutations in a cancer genome as superimposed mutation signatures at varying levels of exposure. (a) Each signature is represented by one of the bar charts, and consists of the relative proportions of different types of mutations in that signature. For example, in Signature 1, C > T mutations make up almost half of the total number of mutations, whereas T > A mutations constitute only about 10% of the total. **(b)** Each of the three signatures contributes a different number of mutations to the actual catalog, represented in the 'pie chart'. In this example, Signature 1 contributes 1,000 mutations, Signature 2 contributes 1,500, and Signature 3 contributes 750. The 1,000 mutations from Signature 1 are allocated according to the bar chart that represents the proportions of different types of mutations in this signature. In this case, Signature 1 would contribute approximately 50% × 1,000 = 500C > T mutations. Signature 2 would contribute approximately 9% × 1,500 = 135C > T mutations. Signature 3 would contribute approximately 10% × 750 = 75C > T mutations. The total number of C > T mutations in the reconstructed catalog would be 500 + 135 + 75 = 710. The reconstruction of the **(d)** actual catalog is approximate, and in this example, the reconstruction does not account for 65 mutations, approximately 2% of the total in the actual mutation catalog - the gray noisy line in **(c)**. This figure is a simplification; in fact, in references [8,75,76], signatures are composed of nucleotide mutations in their trinucleotide contexts, as shown in Figures **3**c,d,h and **4**a-d. The mathematical procedures for approximating observed catalogs from mixtures of trinucleotide signatures are the same, but the trinucleotide context provides far more useful information: for example, the spikes in AA-exposed UTUCs show that the AA-induced A > T mutations tend to occur in a (C|T)AG trinucleotide context. Reproduced from [76] with permission from Elsevier.

had it been present. This would suggest that still other important environmental exposures were not represented among the 7,042 tumors. Because environmental exposures vary widely by geography, it will be important to determine somatic mutation catalogs from a diversity of geographic regions. For example, we previously showed that different genes are mutated in cholangiocarcinomas from different geographical regions and with different etiologies [10,11]. In addition, it is crucially important to have detailed clinical information associated with somatic mutation catalogs. It is possible that the mutagenic exposures responsible for some signatures in previous studies [8] could not be identified because the relevant clinical information was not available. For example, exposures to compounds such as aflatoxins would probably not be captured in clinical records. It is also possible that the mutation signatures of some exposures were not detected because the trinucleotide context and other characteristics of the mutations have not been determined from biochemical studies.

The examples of signatures described above focus on single-nucleotide mutations within trinucleotide contexts as the main distinguishing features of signatures. However, other characteristics of mutation catalogs can also be included as features of mutation signatures and analyzed by NMF [8,76]. For example, strand bias could

be included by considering the two strands separately for each class of mutation in transcribed regions; in this case one would consider C > T on the transcribed strand to be distinct from G > A (the complementary mutation). Other types of mutations, including small insertions and deletions and dinucleotide mutations such as those that occur as a result of UV exposure (CC > TT mutations), can also be included as features of mutation signatures. The framework can also be expanded to consider more bases adjacent to the mutated nucleotide - for example, a pentanucleotide rather than a trinucleotide context. The framework can also be applied to specific regions of the genome. For example, the APOBEC signature (Figure 4a) shows strand bias in exons, but not in introns [76]. Given that both exons and introns are transcribed, the exonic strand bias does not seem to be the result of transcription-coupled repair, and the underlying mechanism remains unknown. However, by distinguishing mutations according to whether they occur in exons or introns, this information could be used to generate a more informative mutation signature. The utility of these possible extensions remains untested, but is likely to increase as additional tumor genomes, which capture about 50 times more mutation information than exomes, are sequenced.

Mutation signatures for surveillance and prevention

Much of cancer is associated with exogenous exposures, and therefore in principle amenable to control by avoidance of those exposures. Examples include tobacco smoke, UV light, and many infectious exposures, such as hepatitis B and C, human papilloma virus, and *Helicobacter pylori* [77-79]. IARC lists 422 known or likely exogenous carcinogens [80]. Indeed, prevention by avoidance of exogenous carcinogenic exposures has been an effective long-term strategy for the control of cancer, with tobacco smoking as the most salient example [49,81]. However, evidence from recent work [8] indicates that many exogenous exposures remain unidentified. Notably, as described earlier, of the 21 mutation signatures identified in [8], 10 lacked any known underlying mutational process or exposure, and over two-thirds of cancer types were affected by signatures due to unknown causes. Furthermore, only three exogenous mutagens were identified: tobacco smoking (12% of all tumors), UV light (5% of all tumors), and temozolomide (0.5% of all tumors), and the cause of Signature 5 (found in 14% of all tumors) is unknown. Some cancers were disproportionately affected by signatures with unknown causes. For example, 89% of HCCs showed Signature 12, and 90% showed Signature 16, both with unknown causes. Conversely, the signatures of some well-known mutagens were not detected (Table 1), suggesting that cancers due to these mutagens were rare or non-existent among the 7,042 tumors studied. This implies that the signatures of many exposures have yet to be captured in sequenced tumor exomes or genomes. Thus, the analysis of mutation signatures in catalogs of somatic mutations from tumors is promising but in its infancy. To realize this promise, we must extend our knowledge in two aspects.

The first is to expand the diversities of tumor types and of their geographical origins. There is already rapid growth in the number of sequenced cancer genomes and their catalogs of somatic mutations. An important advantage of next-generation sequencing in this endeavor is that it is based on an inexpensive, commodity technology, the price of which will continue to drop. In addition, next-generation sequencing provides direct readouts of the mutations that actually occur in tumors. In this context, we note that using whole-exome or whole-genome sequencing to detect mutations (rather than sequencing targeted, cancer-related genes) ensures that most mutations detected are selectively inconsequential passengers. Even though a few somatic mutations in whole-exome or whole-genome sequence are drivers, they are so few that they have negligible influence on the signature. Finally, the large amount of data generated by whole-exome and especially whole-genome sequencing provides optimal statistical power to tease apart the signatures of different mutational processes or exposures.

The second aspect in which we must extend our knowledge consists of establishing connections between specific mutagens and their mutation signatures. This is likely to require experimental exposure of cells or animals to mutagens or their biochemically active metabolites, followed by next-generation sequencing of either clonal populations of exposed cells or of tumors that develop in exposed animals. Sequencing of the exposed genomes will connect specific mutagens to their mutation signatures in far more detail than is currently available. When mutation signatures cannot be found among the signatures of known mutagens, this would suggest the effects of an unknown exposure or mutational process, and point to the need for further epidemiological, toxicological, or biological research.

To our knowledge, there has been little work toward this goal, and our work on the mutation signature of AA and its application to detect AA exposure in HCC is an example [16].

Conclusions and future directions

We envision that the groundbreaking technical advances for detection of signatures in genome- and exome-wide catalogs of somatic mutations from thousands of tumors will enable the assembly of a wide-ranging compendium of mutation signatures from diverse cancer types and multiple geographical regions. This compendium would contain many more whole-genome catalogs of somatic mutations (as opposed to exome catalogs) than are

currently available, and would encompass tumors from many more geographical regions, thus capturing a much wider range of mutagenic exposures. This compendium could be combined with experimental determination of the extended signatures of known and suspected mutagens, including, when necessary, their signatures in different tissues or cell types. Signatures with known causes would represent future opportunities for prevention. Signatures with unknown causes would point to the need for further investigation of exogenous mutagens or endogenous mutation processes.

The first part of this vision, the assembly of a compendium of mutation signatures from ever more cancer genomes, seems certain to happen because of the plummeting cost of sequencing and the many ongoing efforts to sequence tumor genomes. Nevertheless, there are many open questions on how best to deploy NMF or NMF-related procedures to assemble this compendium. For example, what factors determine the power of these procedures to distinguish similar mutation signatures? As the number of genome-wide somatic mutation catalogs increases, will it become worthwhile to include additional information, such as strand bias or pentanucleotide context, in mutation signatures? Fortunately, NMF-related procedures are an active area of machine learning research. For example, enhanced NMF procedures that prefer sparser solutions - solutions in which the mutation catalog of a given tumor is modeled as the mixture of a relatively small number of signatures - have been recently proposed [82-85]. Other proposed enhanced NMF procedures could favor solutions with fewer mutation signatures contributing to each tumor, leading to more interpretable results [85-87].

The second part of the vision, the experimental elucidation of signatures and the investigation of possible causes of signatures with unknown causes, will require concerted effort. There will surely be challenges in understanding the signatures of complex mutagens such as tobacco smoke, and challenges in understanding the differences in the mutagens' metabolisms and mutagenic activity across different tissues and cell types. Nevertheless, in the near term it will be possible to dissect and refine the worldwide repertoire of signatures and to assign some of these signatures to known causes as experimental studies advance. Of course, not all cancer is due to mutagenic exposures, but linking somatic mutation catalogs generated by next-generation sequencing to specific exposures via the mutation signatures of these exposures could substantially reduce the burden of avoidable cancer.

Abbreviations

AA: aristolochic acid; HCC: hepatocellular carcinoma; IARC: International Agency for Research on Cancer; NMF: non-negative matrix factorization, UTUC, upper urinary-tract urothelial cancer; UV: ultraviolet.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Ioana Cutcutache, Weng Khong Lim, and Iain Beehuat Tan for comments on the manuscript.

Author details

[1]Laboratory of Cancer Epigenome, Division of Medical Sciences, National Cancer Centre Singapore, 11 Hospital Drive, Singapore 169610, Singapore. [2]Program in Cancer and Stem Cell Biology, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857, Singapore. [3]Duke-NUS Centre for Computational Biology, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857, Singapore. [4]Division of Cellular and Molecular Research, National Cancer Centre Singapore, 11 Hospital Drive, Singapore 169610, Singapore. [5]Cancer Science Institute of Singapore, National University of Singapore, Centre for Life Sciences, 28 Medical Drive, Singapore 117456, Singapore. [6]Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore.

Published: 31 March 2014

References

1. Pott P: **Cancer Scrot.** In *Chirurgical Observations Relative to the Cataract, the Polypus of the Nose, Cancer of the Scrotum, Different Kinds of Ruptures, and the Mortification of the Toes and Feet.* London: Hawes, Clarke, Collins; 1775.
2. Yamagiwa K, Ichikawa K: **Experimental study of the pathogenesis of carcinoma.** *J Cancer Res* 1918, **3:**1–21.
3. Cook JW, Hewett CL, Hieger I: **The isolation of a cancer-producing hydrocarbon from coal tar. Parts I, II, and II.** *J Chem So (Resumed)* 1933, 395–405.
4. Carrell CJ, Carrell TG, Carrell HL, Prout K, Glusker JP: **Benzo[a]pyrene and its analogues: structural studies of molecular strain.** *Carcinogenesis* 1997, **18:**415–422.
5. Penning TM: *Chemical Carcinogenesis.* New York: Humana Press; 2011.
6. Ikehata H, Ono T: **The mechanisms of UV mutagenesis.** *J Radiat Res* 2011, **52:**115–125.
7. Mortelmans K, Zeiger E: **The Ames *Salmonella*/microsome mutagenicity assay.** *Mutat Res* 2000, **455:**29–60.
8. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, *et al*: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500:**415–421.
9. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete V, Michielin O, Muehlethaler K, Speiser D, Beckmann JS, Xenarios I, Halazonetis TD, Jongeneel CV, Stevenson BJ, Antonarakis SE: **Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma.** *Nat Genet* 2012, **44:**133–139.
10. Ong CK, Subimerb C, Pairojkul C, Wongkham S, Cutcutache I, Yu W, McPherson JR, Allen GE, Ng CC, Wong BH, Myint SS, Rajasegaran V, Heng HL, Gan A, Zang ZJ, Wu Y, Wu J, Lee MH, Huang D, Ong P, Chan-on W, Cao Y, Qian CN, Lim KH, Ooi A, Dykema K, Furge K, Kukongviriyapan V, Sripa B, Wongkham C, *et al*: **Exome sequencing of liver fluke-associated cholangiocarcinoma.** *Nat Genet* 2012, **44:**690–693.
11. Chan-On W, Nairismägi ML, Ong CK, Lim WK, Dima S, Pairojkul C, Lim KH, McPherson JR, Cutcutache I, Heng HL, Ooi L, Chung A, Chow P, Cheow PC, Lee SY, Choo SP, Tan IB, Duda D, Nastase A, Myint SS, Wong BH, Gan A, Rajasegaran V, Ng CC, Nagarajan S, Jusakul A, Zhang S, Vohra P, Yu W, Huang D, *et al*: **Exome sequencing identifies distinct mutational patterns in liver fluke-related and non-infection-related bile duct cancers.** *Nat Genet* 2013, **45:**1474–1478.
12. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, Rajasegaran V, Heng HL, Deng N, Gan A, Lim KH, Ong CK, Huang D, Chin SY, Tan IB, Ng CC, Yu W, Wu Y, Lee M, Wu J, Poh D, Wan WK, Rha SY, So J, Salto-Tellez M, Yeoh KG, Wong WK, Zhu YJ, Futreal PA, Pang B, *et al*: **Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes.** *Nat Genet* 2012, **44:**570–574.
13. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona

LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S, CORGI Consortium; WGS500 Consortium, WGS500 Consortium, Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, *et al*: **Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.** *Nat Genet* 2013, **45:**136–144.

14. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C, Auclair D, Lawrence MS, Stojanov P, Cibulskis K, Choi K, de Waal L, Sharifnia T, Brooks A, Greulich H, Banerji S, Zander T, Seidel D, Leenders F, Ansén S, Ludwig C, Engel-Riedel W, Stoelben E, Wolf J, Goparju C, *et al*: **Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.** *Cell* 2012, **150:**1107–1120.
15. Nagarajan N, Bertrand D, Hillmer AM, Zang ZJ, Yao F, Jacques PE, Teo AS, Cutcutache I, Zhang Z, Lee WH, Sia YY, Gao S, Ariyaratne PN, Ho A, Woo XY, Veeravali L, Ong CK, Deng N, Desai KV, Khor CC, Hibberd ML, Shahab A, Rao J, Wu M, Teh M, Zhu F, Chin SY, Pang B, So JB, Bourque G, *et al*: **Whole-genome reconstruction and mutational signatures in gastric cancer.** *Genome Biol* 2012, **13:**R115.
16. Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, Weng WH, Siew EY, Liu Y, Heng HL, Chong SC, Gan A, Tay ST, Lim WK, Cutcutache I, Huang D, Ler LD, Nairismägi ML, Lee MH, Chang YH, Yu KJ, Chan-On W, Li BK, Yuan YF, Qian CN, Ng KF, Wu CF, Hsu CL, Bunte RM, Stratton MR, *et al*: **Genome-wide mutational signatures of aristolochic acid and its application as a screening tool.** *Sci Transl Med* 2013, **5:**197ra101.
17. Chen CH, Dickman KG, Moriya M, Zavadil J, idorenko VS, Edwards KL, Gnatenko DV, Wu L, Turesky RJ, Wu XR, Pu YS, Grollman AP: **Aristolochic acid-associated urothelial cancer in Taiwan.** *Proc Natl Acad Sci U S A* 2012, **109:**8241–8246.
18. Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R, Turesky RJ, Pu YS, Vogelstein B, Papadopoulos N, Grollman AP, Kinzler KW, Rosenquist TA: **Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing.** *Sci Transl Med* 2013, **5:**197ra102.
19. Hollstein M, Moriya M, Grollman AP, Olivier M: **Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid.** *Mut Res* 2013, **753:**41–49.
20. Gokmen MR, Cosyns JP, Arlt VM, Stiborova M, Phillips DH, Schmeiser HH, Simmonds MS, Cook HT, Vanherweghem JL, Nortier JL, Lord GM: **The epidemiology, diagnosis, and management of aristolochic acid nephropathy: a narrative review.** *Ann Intern Med* 2013, **158:**469–477.
21. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, Ren X, Cibulskis K, Sivachenko AY, Wagle N, Sucker A, Sougnez C, Onofrio R, Ambrogio L, Auclair D, Fennell T, Carter SL, Drier Y, Stojanov P, Singer MA, Voet D, Jing R, Saksena G, Barretina J, *et al*: **Melanoma genome sequencing reveals frequent PREX2 mutations.** *Nature* 2012, **485:**502–506.
22. Drobetsky EA, Grosovsky AJ, Glickman BW: **The specificity of UV-induced mutations at an endogenous locus in mammalian cells.** *Proc Natl Acad Sci U S A* 1987, **84:**9103–9107.
23. Besaratinia A, Li H, Yoon JI, Zheng A, Gao H, Tommasi S: **A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens.** *Nucleic Acids Res* 2012, **40:**e116.
24. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordoñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, *et al*: **A small-cell lung cancer genome with complex signatures of tobacco exposure.** *Nature* 2010, **463:**184–190.
25. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P: **Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.** *Oncogene* 2002, **21:**7435–7451.
26. Phillips DH: **Smoking-related DNA and protein adducts in human tissues.** *Carcinogenesis* 2002, **23:**1979–2004.
27. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD: **Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines.** *Chest* 2013, **143:**e1S–e29S.
28. Hecht SS: **Lung carcinogenesis by tobacco smoke.** *Int J Cancer* 2012, **131:**2724–2732.
29. Ross RK, Yuan JM, Yu MC, Wogan GN, Qian GS, Tu JT, Groopman JD, Gao YT, Henderson BE: **Urinary aflatoxin biomarkers and risk of hepatocellular carcinoma.** *Lancet* 1992, **339:**943–946.
30. Smela ME, Currier SS, Bailey EA, Essigmann JM: **The chemistry and biology of aflatoxin B: from mutational spectrometry to carcinogenesis.** *Carcinogenesis* 2001, **22:**535–545.
31. Bailey EA, Iyer RS, Stone MP, Harris TM, Essigmann JM: **Mutational properties of the primary aflatoxin B1-DNA adduct.** *Proc Natl Acad Sci U S A* 1996, **93:**1535–1539.
32. Shen HM, Ong CN: **Mutations of the p53 tumor suppressor gene and ras oncogenes in aflatoxin hepatocarcinogenesis.** *Mutat Res* 1996, **366:**23–44.
33. Chen CJ, Wang LY, Lu SN, Wu MH, You SL, Zhang YJ, Wang LW, Santella RM: **Elevated aflatoxin exposure and increased risk of hepatocellular carcinoma.** *Hepatology* 1996, **24:**38–42.
34. Bannasch P, Khoshkhou NI, Hacker HJ, Radaeva S, Mrozek M, Zillmann U, Kopp-Schneider A, Haberkorn U, Elgas M, Tolle T, *et al*: **Synergistic hepatocarcinogenic effect of hepadnaviral infection and dietary aflatoxin B1 in woodchucks.** *Cancer Res* 1995, **55:**3318–3330.
35. Wild CP, Montesano R: **A model of interaction: aflatoxins and hepatitis viruses in liver cancer aetiology and prevention.** *Cancer Lett* 2009, **286:**22–28.
36. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, Asthana S, Jalbert LE, Nelson SJ, Bollen AW, Gustafson WC, Charron E, Weiss WA, Smirnov IV, Song JS, Olshen AB, Cha S, Zhao Y, Moore RA, Mungall AJ, Jones SJ, Hirst M, Marra MA, Saito N, Aburatani H, Mukasa A, *et al*: **Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma.** *Science* 2014, **343:**189–193.
37. Yip S, Miao J, Cahill DP, Iafrate AJ, Aldape K, Nutt CL, Louis DN: **MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance.** *Clin Cancer Res* 2009, **15:**4622–4629.
38. Gaskell M, McLuckie KI, Farmer PB: **Comparison of the repair of DNA damage induced by the benzene metabolites hydroquinone and p-benzoquinone: a role for hydroquinone in benzene genotoxicity.** *Carcinogenesis* 2005, **26:**673–680.
39. Saberi Hosnijeh F, Christopher Y, Peeters P, Romieu I, Xun W, Riboli E, Raaschou-Nielsen O, Tjonneland A, Becker N, Nieters A, Trichopoulou A, Bamia C, Orfanos P, Oddone E, Luján-Barroso L, Dorronsoro M, Navarro C, Barricarte A, Molina-Montes E, Wareham N, Vineis P, Vermeulen R: **Occupation and risk of lymphoid and myeloid leukaemia in the European Prospective Investigation into Cancer and Nutrition (EPIC).** *Occup Environ Med* 2013, **70:**464–470.
40. Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R: **Occupational benzene exposure and the risk of chronic myeloid leukemia: a meta-analysis of cohort studies incorporating study quality dimensions.** *Am J Ind Med* 2012, **55:**779–785.
41. Burns MB, Temiz NA, Harris RS: **Evidence for APOBEC3B mutagenesis in multiple human cancers.** *Nat Genet* 2013, **45:**977–983.
42. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, Yee D, Temiz NA, Donohue DE, McDougle RM, Brown WL, Law EK, Harris RS: **APOBEC3B is an enzymatic source of mutation in breast cancer.** *Nature* 2013, **494:**366–370.
43. Vanherweghem JL, Depierreux M, Tielemans C, Abramowicz D, Dratwa M, Jadoul M, Richard C, Vandervelde D, Verbeelen D, Vanhaelen-Fastre R, Vanhaelen M: **Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs.** *Lancet* 1993, **341:**387–391.
44. Bhattacharjee P, Bhattacharyya D: **Characterization of the aqueous extract of the root of Aristolochia indica: evaluation of its traditional use as an antidote for snake bites.** *J Ethnopharmacol* 2013, **145:**220–226.
45. Debelle FD, Vanherweghem JL, Nortier JL: **Aristolochic acid nephropathy: a worldwide problem.** *Kidney Int* 2008, **74:**158–169.
46. Heinrich M, Chan J, Wanke S, Neinhuis C, Simmonds MS: **Local uses of Aristolochia species and content of nephrotoxic aristolochic acid 1 and 2–a global assessment based on bibliographic sources.** *J Ethnopharmacol* 2009, **125:**108–144.
47. Moriya M, Slade N, Brdar B, Medverec Z, Tomic K, Jelakovic B, Wu L, Truong S, Fernandes A, Grollman AP: **TP53 Mutational signature for aristolochic acid: an environmental carcinogen.** *Int J Cancer* 2011, **129:**1532–1536.
48. Arlt VM, Stiborova M, Schmeiser HH: **Aristolochic acid as a probable human cancer hazard in herbal remedies: a review.** *Mutagenesis* 2002, **17:**265–277.
49. US Department of Health and Human Services: *The Health Consequences of Smoking - 50 Years of Progress, A Report of the Surgeon General 2014.* Rockville, MD: Public Health Service Office of the Surgeon General; 2014.

50. Talikka M, Sierro N, Ivanov NV, Chaudhary N, Peck MJ, Hoeng J, Coggins CR, Peitsch MC: **Genomic impact of cigarette smoke, with application to three smoking-related diseases.** *Crit Rev Toxicol* 2012, **42:**877–889.
51. Hecht SS: **Tobacco carcinogens, their biomarkers and tobacco-induced cancer.** *Nat Rev Cancer* 2003, **3:**733–744.
52. Bedard LL, Massey TE: **Aflatoxin B1-induced DNA damage and its repair.** *Cancer Lett* 2006, **241:**174–183.
53. Eaton DL, Gallagher EP: **Mechanisms of aflatoxin carcinogenesis.** *Annu Rev Pharmacol Toxicol* 1994, **34:**135–172.
54. **Aflatoxin.** *IARC Monographs* 1985, 3–87.
55. Ozturk M: **p53 mutation in hepatocellular carcinoma after aflatoxin exposure.** *Lancet* 1991, **338:**1356–1359.
56. Gouas D, Shi H, Hainaut P: **The aflatoxin-induced TP53 mutation at codon 249 (R249S): biomarker of exposure, early detection and target for therapy.** *Cancer Lett* 2009, **286:**29–37.
57. Friedman HS, Kerby T, Calvert H: **Temozolomide and treatment of malignant glioma.** *Clin Cancer Res* 2000, **6:**2585–2597.
58. Payne MJ, Pratap SE, Middleton MR: **Temozolomide in the treatment of solid tumours: current results and rationale for dosing/scheduling.** *Crit Rev Oncol Hematol* 2005, **53:**241–252.
59. Newlands ES, Stevens MF, Wedge SR, Wheelhouse RT, Brock C: **Temozolomide: a review of its discovery, chemical properties, pre-clinical development and clinical trials.** *Cancer Treat Rev* 1997, **23:**35–61.
60. Trivedi RN, Almeida KH, Fornsaglio JL, Schamus S, Sobol RW: **The role of base excision repair in the sensitivity and resistance to temozolomide-mediated cell death.** *Cancer Res* 2005, **65:**6394–6400.
61. Margison GP, Santibanez Koref MF, Povey AC: **Mechanisms of carcinogenicity/chemotherapy by O6-methylguanine.** *Mutagenesis* 2002, **17:**483–487.
62. Caporali S, Falcinelli S, Starace G, Russo MT, Bonmassar E, Jiricny J, D'Atri S: **DNA damage induced by temozolomide signals to both ATM and ATR: role of the mismatch repair system.** *Mol Pharmacol* 2004, **66:**478–491.
63. **A review of human carcinogens: Chemical agents and related occupations.** *IACR Monographs* 2012, **100F**.
64. Cogliano VJ, Baan R, Straif K, IARC Monographs programme staff: **Updating IARC's carcinogenicity assessment of benzene.** *Am J Ind Med* 2011, **54:**165–167.
65. Whysner J, Reddy MV, Ross PM, Mohan M, Lax EA: **Genotoxicity of benzene and its metabolites.** *Mutat Res* 2004, **566:**99–130.
66. Nakayama A, Noguchi Y, Mori T, Morisawa S, Yagi T: **Comparison of mutagenic potentials and mutation spectra of benzene metabolites using supF shuttle vectors in human cells.** *Mutagenesis* 2004, **19:**91–97.
67. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA: **An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.** *Nat Genet* 2013, **45:**970–976.
68. Fishel R, Kolodner RD: **Identification of mismatch repair genes and their role in the development of cancer.** *Curr Opin Genet Dev* 1995, **5:**382–395.
69. Peto J: **Cancer epidemiology in the last century and the next decade.** *Nature* 2001, **411:**390–395.
70. Edwards BK, Howe HL, Ries LA, Thun MJ, Rosenberg HM, Yancik R, Wingo PA, Jemal A, Feigal EG: **Annual report to the nation on the status of cancer, 1973–1999, featuring implications of age and aging on U.S. cancer burden.** *Cancer* 2002, **94:**2766–2792.
71. Frank SA: *Dynamics of Cancer Incidence, Inheritance, and Evolution.* Princeton (NJ): Princeton University Press; 2007.
72. Hoeijmakers JH: **DNA damage, aging, and cancer.** *N Engl J Med* 2009, **361:**1475–1485.
73. Lombard DB, Chua KF, Mostoslavsky R, Franco S, Gostissa M, Alt FW: **DNA repair, genome stability, and aging.** *Cell* 2005, **120:**497–512.
74. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, *et al*: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446:**153–158.
75. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, *et al*: **Mutational processes molding the genomes of 21 breast cancers.** *Cell* 2012, **149:**979–993.
76. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: **Deciphering signatures of mutational processes operative in human cancer.** *Cell Rep* 2013, **3:**246–259.
77. Moudgil V, Redhu D, Dhanda S, Singh J: **A review of molecular mechanisms in the development of hepatocellular carcinoma by aflatoxin and hepatitis B and C viruses.** *J Environ Pathol Toxicol Oncol* 2013, **32:**165–175.
78. Micu G, Staniceanu F, Zurac S, Bastian A, Gramada E, Nichita L, Popp C, Sticlaru L, Andrei R, Socoliuc C: **Carcinogenesis and infection with *Helicobacter pylori*.** *Rom J Intern Med* 2010, **48:**299–306.
79. Deligeoroglou E, Christopoulos P, Aravantinos L, Papadias K: **Human papilloma virus molecular profile and mechanisms of cancerogenesis: a review.** *Eur J Gynaecol Oncol* 2009, **30:**128–132.
80. **Review of human carcinogens.** *IARC Monographs* 2012, **100**.
81. Jemal A, Thun MJ, Ries LA, Howe HL, Weir HK, Center MM, Ward E, Wu XC, Eheman C, Anderson R, Ajani UA, Kohler B, Edwards BK: **Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control.** *J Natl Cancer Inst* 2008, **100:**1672–1694.
82. Guan N, Huang X, Lan L, Luo Z, Zhang X: **Graph based semi-supervised non-negative matrix factorization for document clustering**. In *2012 11th International Conference on Machine Learning and Applications (ICMLA)*, Volume 1. Institute of Electrical and Electronics Engineers; 2012, 404–408.
83. Hillebrand M, Krebel U, Wohler C, Kummert F: **Traffic sign classifier adaption by semi-supervised co-training.** *Artificial Neural Networks Pattern Recognition* 2012:193–200.
84. Lefevre A, Bach F, Fevotte C: **Semi-supervised NMF with time-frequency annotations for single-channel source separation.** *ISMIR 2012: 13th International Society for Music Information Retrieval Conference* [http://ismir2012.ismir.net/event/papers/115-ismir-2012.pdf].
85. Morikawa Y, Yukawa M: **A sparse optimization approach to supervised NMF based on convex analytic method**. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): May 2013*. Vancouver, Canada: Institute of Electrical and Electronics Engineers; 2013:6078–6082.
86. Chen M, Chen WS, Chen B, Pan B: **Non-negative sparse representation based on block NMF for face recognition.** *Lecture Notes Comput Sci* 2013, **8232:**26–33.
87. Sindhwani V, Ghoting A: **Large-scale distributed non-negative sparse coding and sparse dictionary learning**. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: August 2012*. Beijing, China: Association for Computing Machinery; 2012:489–497.

doi:10.1186/gm541
Cite this article as: Poon *et al.*: Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Medicine* 2014 **6**:24.