Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors

Mi Ni Huang,[1,2,9] Willie Yu,[1,2,9] Wei Wei Teoh,[3] Maude Ardin,[4] Apinya Jusakul,[1,5] Alvin Wei Tian Ng,[2,6] Arnoud Boot,[1,2] Behnoush Abedi-Ardekani,[7] Stephanie Villar,[4] Swe Swe Myint,[5] Rashidah Othman,[3] Song Ling Poon,[1,5] Adriana Heguy,[8] Magali Olivier,[4] Monica Hollstein,[4] Patrick Tan,[1] Bin Tean Teh,[1,5] Kanaga Sabapathy,[1,3] Jiri Zavadil,[4] and Steven G. Rozen[1,2]

[1]Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, 169857, Singapore; [2]Centre for Computational Biology, Duke-NUS Medical School, 169857, Singapore; [3]Division of Cellular & Molecular Research, Humphrey Oei Institute of Cancer Research, National Cancer Centre, 169610, Singapore; [4]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, 69008 Lyon, France; [5]Laboratory of Cancer Epigenome, Division of Medical Sciences, National Cancer Centre, 169610, Singapore; [6]NUS Graduate School for Integrative Sciences and Engineering, 117456, Singapore; [7]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, 69008 Lyon, France; [8]Department of Pathology and Genome Technology Center, New York University Langone Medical Center, New York, New York 10016, USA

Aflatoxin B1 (AFB1) is a mutagen and IARC (International Agency for Research on Cancer) Group 1 carcinogen that causes hepatocellular carcinoma (HCC). Here, we present the first whole-genome data on the mutational signatures of AFB1 exposure from a total of >40,000 mutations in four experimental systems: two different human cell lines, in liver tumors in wild-type mice, and in mice that carried a hepatitis B surface antigen transgene—this to model the multiplicative effects of aflatoxin exposure and hepatitis B in causing HCC. AFB1 mutational signatures from all four experimental systems were remarkably similar. We integrated the experimental mutational signatures with data from newly sequenced HCCs from Qidong County, China, a region of well-studied aflatoxin exposure. This indicated that COSMIC mutational signature 24, previously hypothesized to stem from aflatoxin exposure, indeed likely represents AFB1 exposure, possibly combined with other exposures. Among published somatic mutation data, we found evidence of AFB1 exposure in 0.7% of HCCs treated in North America, 1% of HCCs from Japan, but 16% of HCCs from Hong Kong. Thus, aflatoxin exposure apparently remains a substantial public health issue in some areas. This aspect of our study exemplifies the promise of future widespread resequencing of tumor genomes in providing new insights into the contribution of mutagenic exposures to cancer incidence.

[Supplemental material is available for this article.]

Many mutagens and mutagenic processes imprint a characteristic "extended mutational signature" that comprises single nucleotide substitutions in the context of the preceding and following bases, as well as additional features, including transcription and replication strand bias, association with indels or dinucleotide substitutions, and so forth. For example, ultraviolet light tends to induce TCC > TTC mutations across the entire genome, and aristolochic acid induces CAG > CTG mutations (Alexandrov et al. 2013; Hoang et al. 2013; Poon et al. 2013). Both show transcriptional-strand bias but negligible replication strand bias, and neither is associated with elevated levels of indel mutations. Unlike aristolochic acid, however, UV light induces high levels of CC > TT dinucleotide substitutions. Elucidation of mutational signatures has been driven by computational analyses of somatic mutations in thousands of tumors, followed by integration with clinical information and prior knowledge of mutational mechanisms (Alexandrov et al. 2013). There are now 30 widely accepted signatures, with a variety of known, suspected, or unknown causes (http://cancer.sanger.ac.uk/cosmic/signatures). These signatures hold promise in molecular cancer epidemiology because they bear witness to mutagenic exposures and because they illuminate endogenous mutagenic processes and mechanisms of DNA damage and repair (Alexandrov and Stratton 2014; Helleday et al. 2014; Poon et al. 2014; Nik-Zainal et al. 2015; Hollstein et al. 2016).

Despite the progress in computational analysis of mutational signatures, the experimental delineation of extended mutational signatures across whole exomes or whole genomes is essential because the causes of many computationally discerned signatures are unknown and because it is possible that several mutagenic processes or exposures could generate similar signatures. While exome- or genome-wide extended mutational signatures from several metazoan experimental systems have been reported (Meier et al. 2014; Olivier et al. 2014; Severson et al. 2014; Nik-Zainal et al. 2015; Poon et al. 2015; Blokzijl et al. 2016; Zamborszky et al. 2016), development of systems that robustly recapitulate in vivo human mutagenesis remains a challenge.

[9]Co-first authors
Corresponding authors: cmrksb@nccs.com.sg, zavadilj@iarc.fr, steve.rozen@duke-nus.edu.sg

To study the connections between experimentally and computationally derived signatures, we focused on the mutagenic carcinogen aflatoxin B1 (AFB1). AFB1 is considered the most important of several co-occurring aflatoxins that are produced by molds growing in grain, peanuts, or other food. Aflatoxin exposure is thought to be a major public health risk in parts of Africa and Asia, partly by synergizing with hepatitis B to dramatically increase risk of hepatocellular carcinoma (HCC) (Kensler et al. 2003; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2012). However, to our knowledge, the extended mutational signatures of aflatoxins have been little-studied in the context of normal metazoan DNA repair: a study in *Caenorhabditis elegans* examined the AFB1 mutation signature based on 31 single base-substitution mutations (Meier et al. 2014), and ultradeep sequencing of 6.4 kilobases (kb) in AFB1-exposed mice inferred an AFB1 signature from 397 mutations (Chawanthayatham et al. 2017).

Conversely, a computationally extracted signature of aflatoxin exposure in HCCs, termed COSMIC Signature 24, has been described based on computational analysis of somatic mutations from the whole-exome sequencing (WES) of 11 tumors for which aflatoxin exposure was inferred from African or Asian origins or "race" and, in some HCCs, the presence of the *TP53* R249S mutation, which occurs in roughly half of aflatoxin-exposed HCCs (Hollstein et al. 1991; Montesano et al. 1997; Ming et al. 2002; Sun et al. 2011; Schulze et al. 2015). In addition, while the current paper was in review, a publication on HCCs from Qidong County, a region of known aflatoxin exposure, extracted a signature similar to COSMIC Signature 24 from HCC somatic mutations (Zhang et al. 2017).

Here, we report a first-of-its-kind multisystem strategy for the experimental characterization of the extended mutational signature of AFB1 in human cell lines and in mouse models of AFB1-induced HCC. We further examine the relationships between experimentally determined signatures and mutational signatures in multiple human HCCs with likely dominant aflatoxin exposure and provide a broader overview of the prevalence of likely aflatoxin exposure in published whole-genome sequence (WGS) data from human HCCs.

Results

Mutation spectra of AFB1-exposed human cell lines

To investigate the extended mutational signature of AFB1, we first whole-genome-sequenced two AFB1-exposed HCC cell lines, HepaRG and HepG2 (Figs. 1, 2; Supplemental Table S1). Mutation patterns were remarkably stable across three replicates within each cell line (average pairwise Pearson correlation > 0.97) (Supplemental Table S2; Supplemental Fig. S1). Mutation patterns were also similar between HepG2 and HepaRG but slightly less than within each cell line (average pairwise Pearson correlation = 0.94).

G > T mutations predominated, accounting for 68% and 50% of the total G > N mutations for HepaRG and HepG2, respec-
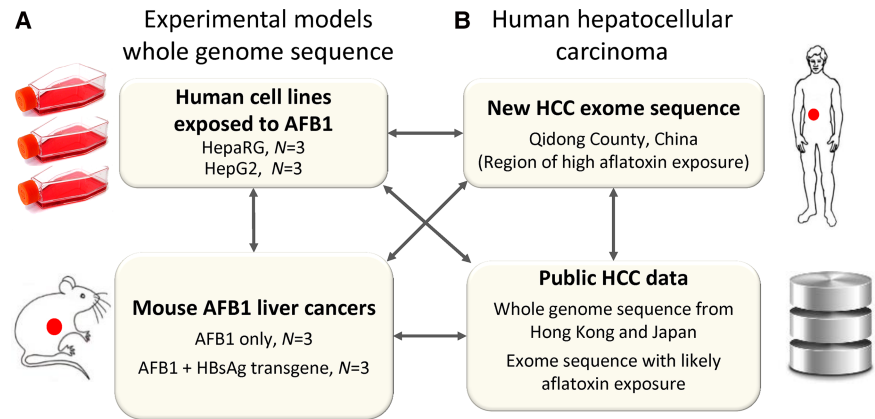

Figure 1. Study design. (*A*) We experimentally elucidated the mutational signature of AFB1 based on whole-genome sequence in four experimental systems: two different AFB1-exposed human cell lines and liver tumors in AFB1-exposed wild-type mice and in AFB1-exposed mice that carried a hepatitis B surface antigen transgene. In total, we examined 48,000 mutations in experimental systems including mutations in cell lines and somatic mutations in mouse tumors. (*B*) We integrated these experimental results with newly generated genomic HCC data from a geographical region of known aflatoxin exposure and with additional, publicly available human HCC data.

tively (Supplemental Table S3). G > A mutations were the next most abundant, followed by G > C mutations. The predominance of G > N mutations and the relative proportions of G > T, G > A, and G > C mutations were broadly consistent with previous experimental studies (Foster et al. 1983; Bailey et al. 1996). In the extended mutational signatures, TGC > TTC mutations were most frequent among the G > T mutations, and AGC > ATC and TGG > TTG mutations were also noticeably enriched (Fig. 2; Supplemental Fig. S1). There was strong enrichment of G > T mutations on the nontranscribed strands of genes (Fig. 2B; Supplemental Fig. S2). The primary mutagenic mechanism of AFB1 is formation of adducts on guanines (Guengerich et al. 1998), and the relative paucity of mutations on the nontranscribed strands is presumably due to transcription-coupled nucleotide-excision repair (TC-NER) of damage occurring on guanines (Fousteri and Mullenders 2008). Furthermore, consistent with TC-NER activity, the level of bias was correlated with mRNA levels in the respective cell lines: in HepaRG only 14% of the mutations in the most highly expressed genes were on the transcribed strand versus 47% in the least expressed genes (Supplemental Table S4). In HepG2, the corresponding figures were 11% and 53%. There is also prior evidence that the effectiveness of TC-NER decreases from the 5′ to 3′ ends of transcripts (Conaway and Conaway 1999; Hu et al. 2015). Consistent with this, and further supporting the involvement of TC-NER in the repair of AFB1-guanine adducts, the ratio of nontranscribed-to transcribed-strand G > T mutations decreased from 2.6:1 in the first 100 kb of transcripts to approximately 1.5:1 in the 200 kb centered on 0.5 Mb ($P < 7 \times 10^{-8}$ for HepaRG and $P < 1.6 \times 10^{-4}$ for HepG2, by logistic regression) (Supplemental Fig. S3).

We also looked for variation in mutation intensity across the genome as a function of replication time or replication strand. In HepG2, but not HepaRG, there was significant enrichment for mutations in late-replicating regions (>56% of G > T mutations in late regions, $P < 6 \times 10^{-6}$ in any HepG2 replicate) (Supplemental Table S5). We are unable to explain the lack of replication-time bias in the HepaRG findings. However, in both cell lines, we observed slight but consistent enrichment of G > T mutations on the leading strand (52.2%, $P < 0.0004$ for HepaRG and 51.6%, $P < 0.032$ for HepG2) (Supplemental Table S6).

We also analyzed the AFB1 mutation data for dinucleotide mutations from guanines, that is, substitutions of two nucleotides: GN > NN or NG > NN. These "G dinucleotide mutations" are associated with COSMIC Signature 4, which is associated with tobacco smoking and also has a high proportion of G > T mutations (Alexandrov et al. 2013). In the AFB1-exposed cells, the ratios of the number of G dinucleotide mutations to G > N mutations were <0.011 (Supplemental Table S3), lower than any of the HCCs with COSMIC Signature 4 in Fujimoto et al. (2016) (minimum 0.0134).

Mutation spectra of mouse AFB1 HCC-like liver tumors

We studied mutation spectra in two mouse models of aflatoxin-associated HCCs. One model consisted of tumors from mice exposed only to AFB1. The other model consisted of tumors from mice bearing the weakly oncogenic hepatitis B surface antigen (HBsAg) transgene (Chisari et al. 1985) and exposed to AFB1, a system designed to model the synergistic carcinogenic effects of hepatitis infection and aflatoxin exposure (Kensler et al. 2003; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans 2012). The WGS-based mutation spectra of all the mouse tumors, like the human cell lines, were dominated by G > T mutations and also showed noticeably high numbers of G > C and G > A mutations (Fig. 3A,B). However, AFB1+HBsAg tumors had far fewer

G > N mutations (mean 468) than the AFB1-only tumors (mean 3242).

We observed more variability between spectra among the mouse tumors than among each cell line, with an average Pearson correlation of 0.89 among the AFB1-only tumors, compared to 0.97 within each cell line and 0.94 among all cell lines (Supplemental Table S2). This was also apparent in a principal components analysis (PCA) (Fig. 3D). The AFB1+HBsAg tumors had even more diversity, probably due primarily to the greater relative sampling variation implicit in fewer mutations. Additional variability among the mouse tumors may have been due to additional mutagenic processes before and during cancer development. Notably, in addition to the G > N mutations attributable to AFB1, the spectra from M1, M2, and M6 contained COSMIC Signature 17. This signature, which is of unknown etiology, is characterized by isolated peaks at AAN > ACN, especially at AAG > ACG, and at AAG > AGG (Alexandrov et al. 2013). We considered the possibility that the emergence of COSMIC Signature 17 in these tumors might stem from some somatic mutation affecting DNA repair or activating some endogenous mutagenic process. However, we were not able to identify any somatic mutations in the affected tumors that might account for this additional signature.

To further investigate sources of diversity among the mutation spectra, we stratified the spectra by variant allele frequency (VAF), as variants with lower VAF are often subclonal and, as
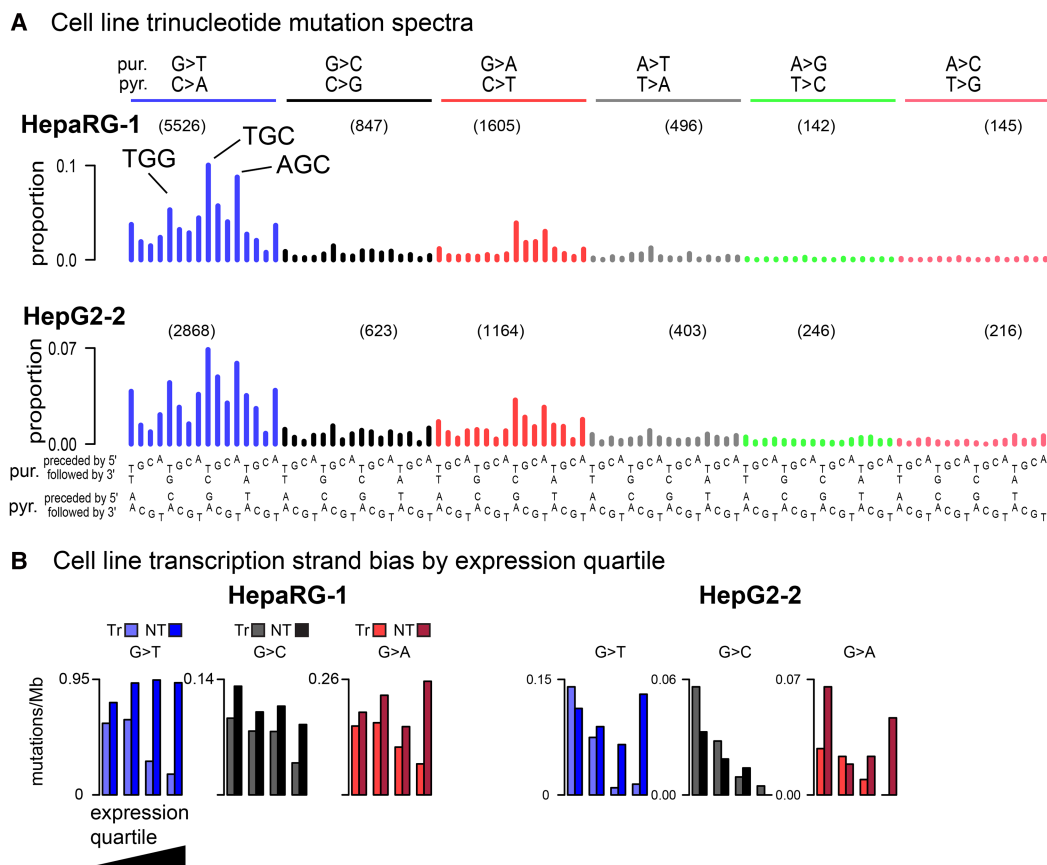

Figure 2. (A) Representative human-cell line trinucleotide mutation spectra grouped by mutations from guanine (G > T, G > C, G > A) and adenine (e.g., A > T, A > G, A > C). The most frequent G > T mutations are indicated (TGG > TTG, TGC > TTC, AGC > ATC). The number of mutations in in each mutation class (e.g., G > T) are indicated in parentheses. As there was little variation between replicates within each cell line, we show all individual spectra in Supplemental Figure S1. (B) Extreme transcription-strand bias for genes with high expression levels; see Supplemental Figure S2 for transcription-strand bias for all cell line replicates. Transcribed strand (Tr); nontranscribed strand (NT).
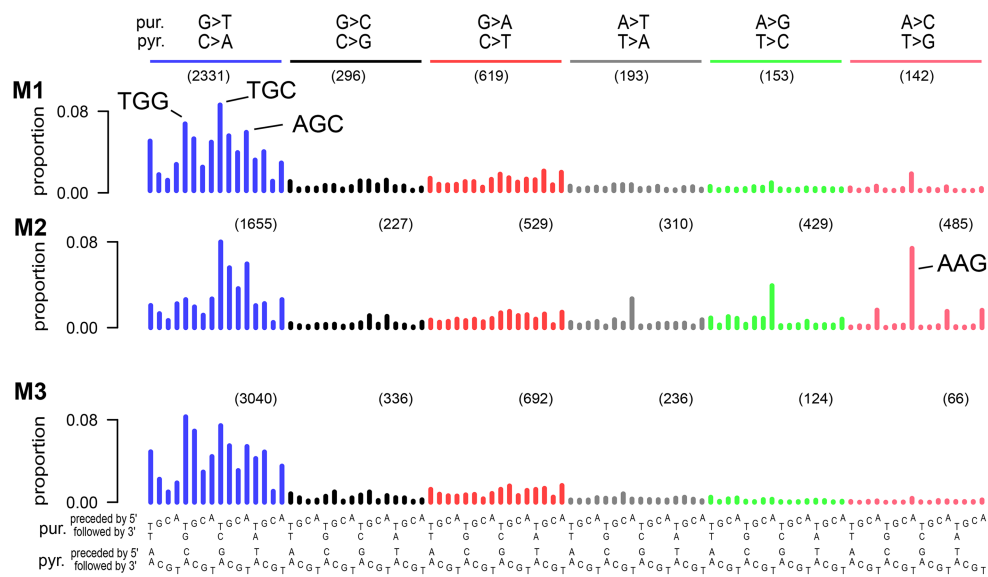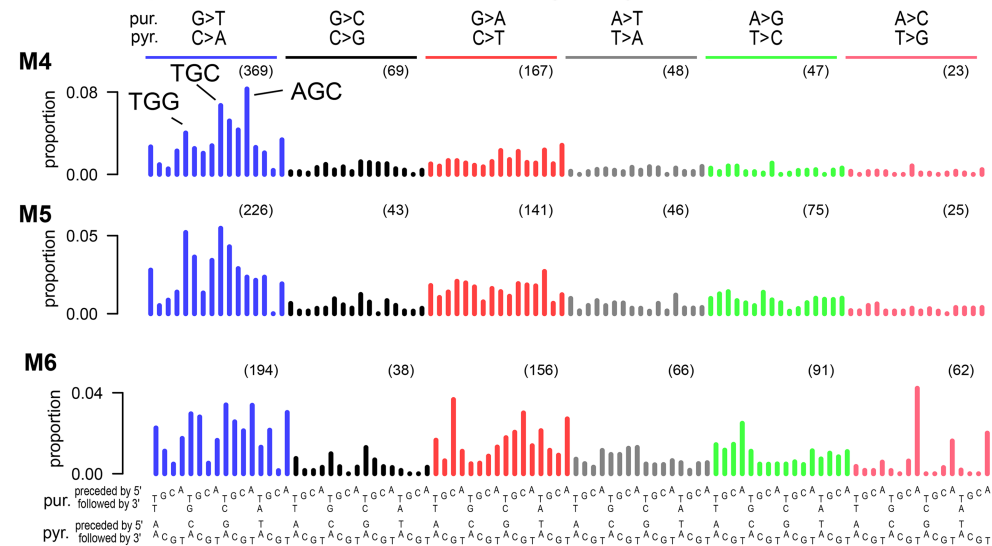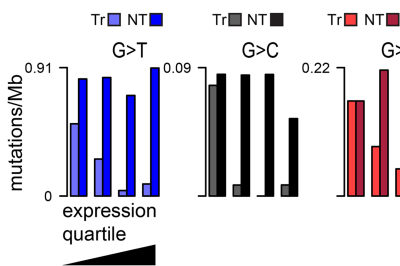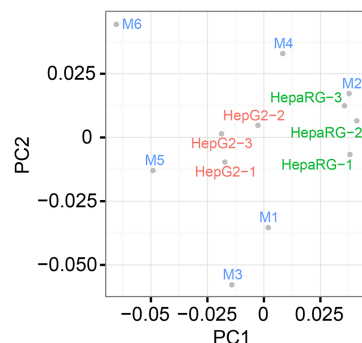
Figure 3. Somatic mutation spectra from HCC-like liver tumors from (*A*) three AFB1-exposed mice and (*B*) three AFB1-exposed mice with an HBsAg transgene. The latter have only one-tenth as many mutations as tumors from the AFB1-only mice. Spectra in panels *A* and *B* were normalized to the trinucleotide frequencies in the human genome. (*C*) Extreme transcription-strand bias for all G > N mutations in highly expressed genes in mouse M1. See Supplemental Figure S3 for other mice. Because of low mutation count, transcription-strand bias is evident only in the G > T mutations in the tumors from the AFB1 + HBsAg mice. Transcribed strand (Tr); nontranscribed strand (NT). (*D*) Principal components analysis (PCA) on G > N mutations in trinucleotide context. Replicates of each of the cell lines cluster together, while the mouse tumors are more dispersed in principal components space. The greater dispersion among the HBsAg tumors (M4, M5, M6) is likely due to higher stochastic variance because of much lower mutation counts combined with greater relative contributions from other mutational processes that arose during tumor development.

such, are likely to have arisen later in tumor development. We also analyzed the tumors for aneuploidy and chromosomal copy number to allow better inference of clonality from the VAFs (Supplemental Figs. S4–S15). AFB1 was administered in a single dose at postnatal day 7, leading to the expectation that all or most AFB1-associated mutations would have occurred prior to tumor initiation and would be clonal. Consistent with this, the analysis of the VAFs of G > T mutations indicates that almost all are clonal. Interestingly, the COSMIC Signature 17 mutations were subclonal and therefore presumably occurred later in tumor development. There was also a diffuse signature similar to COSMIC Signature 5 noticeable in mutations with VAFs < 0.2 in M3 (Supplemental Fig. S8).

Despite the greater diversity in mouse tumor G > N spectra, they broadly resembled the cell lines. As in the cell lines, TGC > TTC, AGC > ATC, and TGG > TTG mutations were prominent in all AFB1-only tumors and in the much less mutated AFB1 +HBsAg tumors. In the AFB1-only tumors, G > T mutations constituted 72% of the total G > N mutations on average (Supplemental Table S3), which was similar to the proportion in the HepaRG cells. In the AFB1+HBsAg tumors, G > T mutations constituted 55% of all G > N mutations, possibly because of G > C and G > A mutations due to factors other than AFB1 exposure. As in the cell lines, the ratio of G dinucleotide mutations to the total number of G > N mutations was low (<0.0123 in all tumors) (Supplemental Table S3).

Also like the human cell lines, and consistent with the operation of TC-NER on guanine adducts, the mouse tumors were strongly enriched for G > T mutations on the nontranscribed strands of genes, and the level of enrichment was higher in more highly expressed genes: among the AFB1-only tumors, an average of only 7% of G > T mutations occurred on the transcribed strands of genes in the top expression quartile, and this pattern of expression-level-correlated strand bias extended to G > C and G > A mutations (Fig. 3C; Supplemental Fig. S16; Supplemental Table S4). Among the AFB1+HBsAg tumors, G > T strand bias in the top expression quartile was also detectable in the aggregate, with 26% of these mutations on the transcribed strand. We hypothesize that the weaker strand bias reflects the presence of other mutational processes that also generate G:C > T:A mutations, which would have had more relative impact in the low-mutation AFB1+HBsAg tumors. Also like the human cell lines, and consistent with the operation of TC-NER of guanine adducts, transcription-strand bias decreased with distance from the 5′ end of the gene: the ratio of nontranscribed to transcribed-strand G > T mutations decreased from 4.1:1 in the first 100 kb to 1.5:1 at the 200 kb centered at 0.5 Mb ($P < 1.7 \times 10^{-8}$ by logistic regression) (Supplemental Fig. S17).

We were unable to assess whether there was a difference in mutation intensity between early and late replicating regions or between leading and lagging replication strands, as we could not locate replication timing data for mouse hepatocytes.

Mutation spectra of likely aflatoxin-associated human HCCs

We integrated the experimental results from cell lines and mice with somatic-mutation spectra from newly sequenced HCCs from Qidong, China, a region of well-studied high aflatoxin exposure (Szymanska et al. 2009; Kensler et al. 2011). The HCCs from Qidong were also selected for somatic *TP53* R249S mutations, which is considered indicative of AFB1 exposure (Hollstein et al. 1991; Montesano et al. 1997; Sun et al. 2011). The Qidong HCCs had high proportions of G > T mutations and transcription-strand bias skewed toward highly expressed genes, consistent with afla-

toxin exposure (Supplemental Figs. S18, S19; Supplemental Table S4). In this exome data, there were too few mutations to assess whether transcription-strand bias decreased along the length of transcripts (Supplemental Fig. S20).

Of note, one HCC from Qidong (PT4) (Supplemental Fig. S18) had a prominent A > T mutation pattern indicative of exposure to aristolochic acid (Hoang et al. 2013; Poon et al. 2013). However, considered apart from A > T mutations, the G > N pattern in this HCC was consistent with aflatoxin exposure, suggesting past coexposure to these two strong carcinogens.

We next analyzed additional, publicly available HCC spectra, some with previously hypothesized aflatoxin exposure (Supplemental Table S7). PCA of the proportions of G > T mutations in trinucleotide context separated the experimental spectra from the vast majority of the HCC spectra by lower values of principal component 1 (PC1) (Fig. 4). To select likely aflatoxin-affected HCCs for further study, we used as a threshold the maximum value of PC1 across the experimental data (M3's PC1 value, indicated by a vertical green line) (Fig. 4). We selected the eight WGS HCCs that exceeded this threshold for initial study. Among these eight was DO23048, which was identified in Schulze et al. (2015) from WES data as probably aflatoxin-exposed. All six of the WES HCCs from Qidong exceeded this threshold. Furthermore, the 10 presumed aflatoxin-related WES tumors in Schulze et al. (2015) also exceeded this threshold and showed patterns of G > T and G > N mutations and transcriptional-strand bias similar to that observed in the Qidong HCCs (Supplemental Figs. S21–S24). An additional PCA, based on the proportions of G > N mutations in trinucleotide context, likewise placed the eight selected aflatoxin-linked WGS HCCs and the WES HCCs from Qidong or identified in Schulze et al. (2015) near the experimental samples and away from the vast majority of the remaining HCCs (Supplemental Fig. S25). To gain insight into possible aflatoxin exposure in North America, we repeated the PCA after adding the full set of HCC WES somatic mutations from The Cancer Genome Atlas (TCGA) (https ://cancergenome.nih.gov/). This flagged five additional possible aflatoxin-exposed HCCs. However, examination of the spectra suggested that they did not reflect detectable aflatoxin exposure, because of minimal transcriptional-strand bias and because of the patterns of G > T mutations in trinucleotide context (Supplemental Fig. S26). Thus, as far as can be determined from the TCGA WES somatic mutations, the rate of aflatoxin exposure in HCCs treated in North America was two in 289, or 0.7%.

As in the experimental systems, the spectra of the eight WGS HCCs initially selected for analysis had abundant G > T mutations, and consistent with the operation of TC-NER, marked transcription-strand bias skewed toward highly expressed genes (Fig. 5A; Supplemental Figs. S27, S28; Supplemental Table S4). As was the case for the cell lines and mouse tumors, and again consistent with the operation of TC-NER, transcription-strand bias decreased with distance from the 5′ end of the gene: the ratio of nontranscribed- to transcribed-strand G > T mutations decreased from 2.3:1 in the first 100 kb to 1.3:1 at the 200 kb centered at 0.5 Mb ($P < 3 \times 10^{-10}$ by logistic regression) (Supplemental Fig. S29). In addition, like the HepG2 cells, all eight WGS HCCs showed an excess of likely AFB1-associated mutations in late replicating regions of the genome (>60% in late rather than early replicating regions, $P < 10^{-12}$ for G > T mutations) (Supplemental Table S5). Also, like the cell lines, the eight WGS HCCs had slight but consistent excesses of G > T mutations on leading replication strands (52% of G > T mutations on the leading strand, $P = 1.6 \times 10^{-7}$ by binomial test on aggregated mutation counts) (Supplemental Table S6).
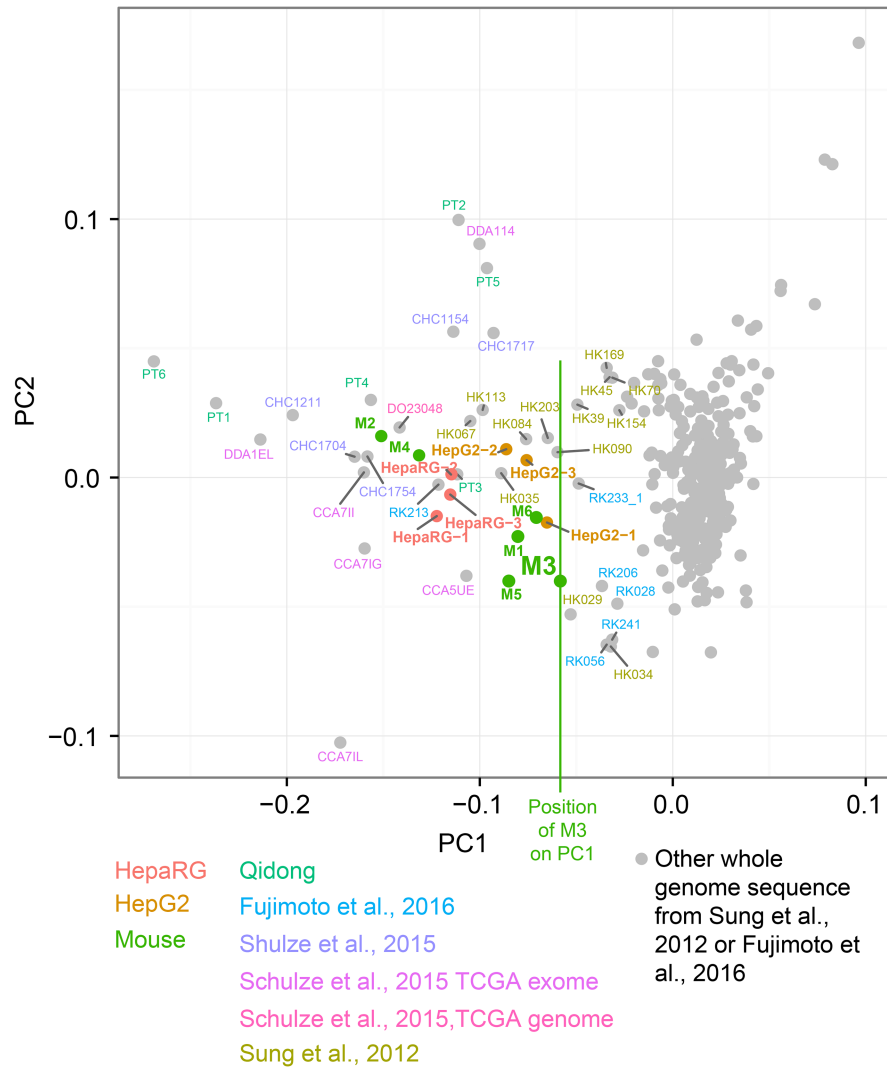
Figure 4. Selection of likely aflatoxin-associated HCCs by principal components analysis on the proportions of G > T mutations in trinucleotide context. Vertical green line indicates the value of M3 in PC1, which was used as a threshold for selecting WGS HCCs likely exposed to aflatoxins for further study: One HCC identified from WES data in Schulze et al. (2015) for which WGS data was subsequently available (DO23048), six from Sung et al. (2012), and one HCC from Fujimoto et al. (2016). As expected due to the higher relative sampling variance in WES spectra due to small numbers of mutations, these were more variable than WGS spectra.

mutations from other processes, especially among the G:C > A:T mutations, which can be generated by several other mutational processes that operate in HCCs (http://cancer.sanger.ac.uk/cosmic/signatures). Indeed, inspection of the spectra suggests presence of COSMIC Signatures 12 and 16, which are both common in HCC and generate G:C > A:T mutations.

Decomposition of aflatoxin signatures and prevalence of aflatoxin exposure in HCC

In view of the differences in the G > T spectra between the aflatoxin associated HCCs and all four experimental systems—namely the predominance of GGC > GTC in HCCs rather than TGC > TTC (Figs. 2A, 3A,B, 5A)—we used a non-negative matrix factorization (NMF) technique to factor the spectra of the AFB1-exposed cell lines and mouse tumors and likely aflatoxin-exposed WGS HCCs into two signatures based on the 48 G > N mutation classes; we call these two signatures $AFB1sig_{G>N}$ and $AFsig2_{G>N}$ (Fig. 6A–C). Reconstruction of the observed G > N spectra was quite accurate (Supplemental Table S8). Given that COSMIC Signatures 5 and 6 have been observed in human HCCs, we sought to determine if they were affecting signature decomposition. To this end, we performed two separate factorizations into three signatures. One factorization specified inclusion of COSMIC Signature 5 and the other specified inclusion of COSMIC 6. Neither signature affected $AFB1sig_{G>N}$, but $AFsig2_{G>N}$ was slightly affected (Supplemental Table S9). It is possible that $AFsig2_{G>N}$ represents simply a subtraction of $AFB1sig_{G>N}$ from the G > N spectra of HCCs rather than an independent mutational process. Clarification of this will depend on future investigation into the causes of differences between the experimental AFB1 spectra and the spectra in HCCs.

We then used the NMF framework and the experimentally derived AFB1-related signatures to analyze the 79 WGS somatic mutation spectra from Sung et al. (2012) and the 268 WGS spectra from Fujimoto et al. (2016) to identify additional HCCs with likely aflatoxin exposure. We analyzed each HCC using COSMIC signatures known to be associated with HCCs but replacing COSMIC Signature 24 with extensions of $AFB1sig_{G>N}$ and $AFsig2_{G>N}$ to all 96 trinucleotide contexts (with all A > N mutation proportions set to 0). We termed these AFB1sig and AFsig2. We did not include COSMIC Signature 24, because we reasoned that both it and the combination of AFB1sig and AFsig2 captured most aspects of aflatoxin exposure, and we sought to assess the presence of AFB1sig and AFsig2 without interference from the largely overlapping

However, the G > T mutations in the HCCs differed from the experimental systems in that the most frequent mutation was GGC > GTC rather than TGC > TTC (Figs. 2A, 3A,B). Consistent with this, PCAs that included only the experimental systems and the likely aflatoxin-exposed HCCs separated these two groups of samples (Fig. 5B,C). We note that the G > N mutation spectra of the eight selected WGS HCCs were more similar to each other as a group than were the AFB1-only mouse tumors (average pairwise Pearson correlation among the WGS HCCs was 0.92; among the AFB1-only mouse tumors, 0.89). Thus, the G > N mutation spectra of these WGS HCCs may reflect a reasonably uniform underlying set of mutational processes. Among the eight selected HCCs, G > T mutations accounted for an average of 58% of G > N mutations (Supplemental Table S3), somewhat less than for cell line and AFB1-only mouse tumors. The relatively higher proportions of non-G > T mutations probably reflects the presence of

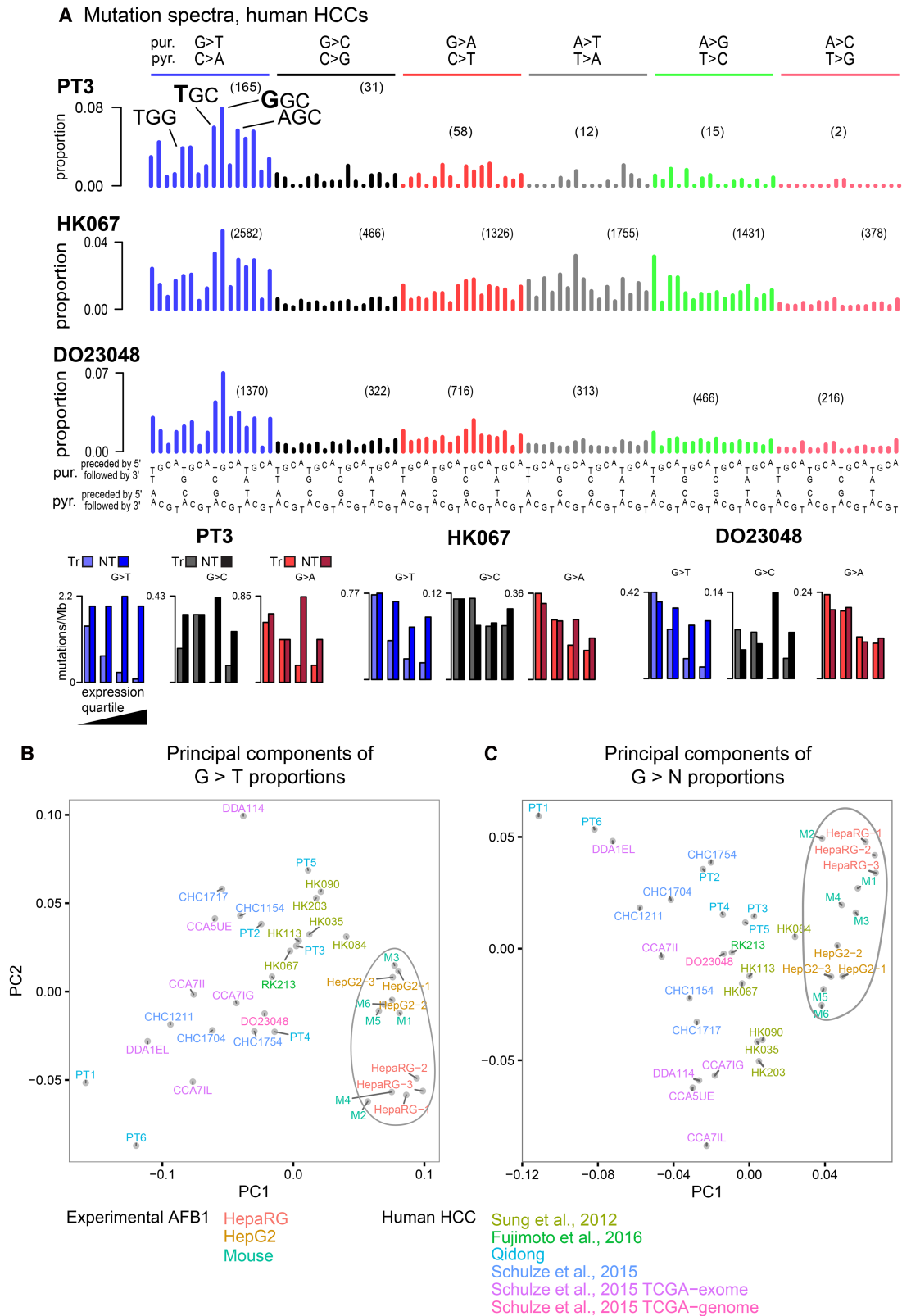**Figure 5.** (*A*) Somatic mutation spectra and transcriptional-strand bias for initial study of aflatoxin signatures in example human HCCs with likely aflatoxin exposure. Transcribed strand (Tr); nontranscribed strand (NT). Principal component analysis of (*B*) G > T and (*C*) G > N mutations for AFB1 experimental data (enclosed in irregular ovals) and HCCs with likely high aflatoxin exposure.
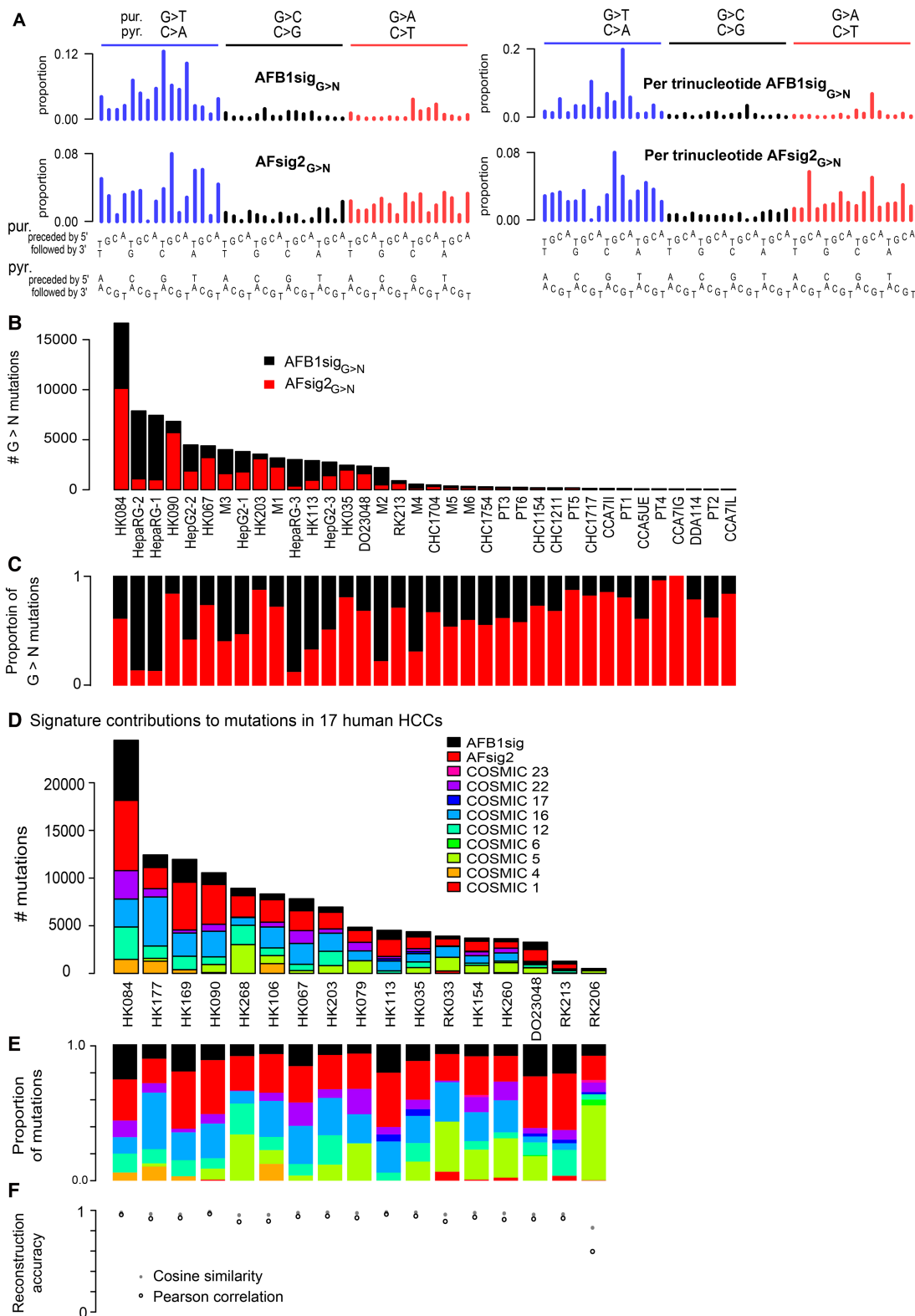
Figure 6. Aflatoxin signatures in human HCCs. (*A*) NMF decomposes the G > N spectra from experimental AFB1 exposure and likely aflatoxin-associated HCCs into two components, denoted AFB1sig$_{G>N}$ and AFsig2$_{G>N}$. The *left*-hand side of the panel shows the signatures, as is conventional, based on trinucleotide frequencies in the human genome; the *right*-hand side shows them in frequencies per trinucleotide—equivalent to assuming all trinucleotides are equally common. (*B,C*) AFB1sig$_{G>N}$ (black) almost completely captures the spectra of HepaRG cell lines; AFsig2$_{G>N}$ (red) almost completely captures the spectra of some HCCs (e.g., HK067, HK203), while the mutations in HepG2, the mouse tumors, and some HCCs are most accurately reconstructed by a mixture of AFB1sig$_{G>N}$ and AFsig2$_{G>N}$. (*D,E*) Nonnegative matrix factorization of selected HCCs using mutational signatures known to occur in HCCs, plus extensions of AFB1sig$_{G>N}$ and AFsig2$_{G>N}$ with A > N proportions set to 0, the latter denoted AFB1sig and AFsig2. (*D*) Absolute mutation numbers assigned to each signature. (*E*) Proportions of mutations assigned to each signature. (*F*) Reconstruction accuracy is generally good, with the exception of RK206, which has few mutations and a hard-to-reconstruct mutation spectrum outside of G > T mutations, partly due to spikes at TCA:TGA > TTA:TAA and CTG:CAG > CCG:CGG.

effects of COSMIC Signature 24. Based on this analysis, in addition to the eight HCCs selected above in Figure 4, nine had substantial evidence for the presence of AFB1sig (i.e., >5% of the mutations attributed to AFB1sig) (Figure 6D–F; Supplemental Table S10). In total, at least 13 of 79 HCCs from Hong Kong (16%) (Sung et al. 2012) and three of the 286 from Japan (1.0%) (Fujimoto et al. 2016) appear to have been exposed to AFB1.

Discussion

This study combined analysis of experimentally induced AFB1 mutational signatures with computational analysis of mutational signatures in human HCCs thought to have been exposed to aflatoxins. The mutational signatures from all four experimental systems were remarkably similar. NMF and PCA indicate that the six *TP53* R249S and hepatitis B-positive HCCs from Qidong were indeed likely exposed to AFB1. NMF and PCA also showed that HCCs in which COSMIC Signature 24 was previously observed (Schulze et al. 2015) were likely exposed to AFB1. By extension, COSMIC Signature 24 likely represents, at least in part, AFB1 exposure. The evidence for these assertions is as follows: (1) The G > N spectra of the likely aflatoxin-related HCCs are similar to the experimental spectra in terms of the proportions of G > T, G > A, and G > C mutations. Furthermore, in PCAs on G > T and G > N mutations (Fig. 4; Supplemental Fig. S13), these HCCs clustered toward the AFB1 experimental systems and away from the overwhelming majority of other HCCs. (2) Both the HCCs and the AFB1 experimental systems showed notable transcription-strand bias that strongly skewed toward highly expressed genes and that declined from the 5′ to 3′ ends of transcripts. (3) The HCCs and the HepG2 cells were enriched for mutations in late replicating regions, and the HCCs and both cell lines were enriched for mutations on the leading replication strand. (We were unable to assess these characteristics in the liver tumors of AFB1-exposed mice.) (4) Both the HCCs and experimental AFB1 systems had low burdens of dinucleotide mutations, lower than any HCC with the smoking-associated COSMIC Signature 4, which might otherwise be difficult to distinguish from aflatoxin exposure. (5) Association of the G > N spectra of the tumors in Schulze et al. (2015) with aflatoxin was supported with geographical information making exposure plausible. This was also true for the six HCCs from Qidong, a region of well-documented aflatoxin exposure. Furthermore, 13 additional HCCs with AFB1sig were from southern China, another region with documented aflatoxin exposure (Wang and Liu 2007). Conversely, only three of 268 HCCs from Japan, a region of low exposure, had AFB1sig. (6) Many of the HCCs with strong evidence of aflatoxin exposure also bore the *TP53* R249S mutation, which sometimes indicates aflatoxin exposure in HCCs (Hollstein et al. 1991; Montesano et al. 1997; Sun et al. 2011). The Qidong HCCs were selected based on the presence of this mutation, while six of the 11 tumors identified in Schulze et al. (2015) had this mutation, as did four of the 12 from Sung et al. (2012) (Supplemental Table S10).

The prevalence of the *TP53* R249S mutation in HCCs with combined hepatitis B and aflatoxin exposure is remarkable, but the sensitivity of R249S as an indicator of aflatoxin exposure has been difficult to study systematically. Partly, this is because peripheral biochemical markers of aflatoxin exposure, such as aflatoxin-albumin or urinary aflatoxin-DNA adducts reflect recent exposure. This makes it difficult to assess the strength of the association between the R249S mutation and long-term aflatoxin exposure (Groopman et al. 1985; Wild et al. 1986; Egner et al. 2006).

Estimates of the proportion of tumors with R249S mutations in HCCs with hepatitis B and aflatoxin exposure have varied by study and may depend on the level of aflatoxin exposure (Hollstein et al. 1991; Montesano et al. 1997; Ming et al. 2002; Szymanska et al. 2009; Sun et al. 2011; Chittmittrapap et al. 2013; Kew 2013). Thus, a detection method based on robust, well-defined mutational signatures would be of interest for identifying HCCs with aflatoxin exposure but lacking the R249S mutation. The WGS-based approach presented here is a step in that direction, promising specificity, if not sensitivity. Given that other exposures, including tobacco smoking and oxidative damage of guanines, also affect the pattern of G > T mutations and could obscure evidence of weak AFB1 exposure, further work in this area is needed.

An unexpected finding in the present study was that the AFB1 experimental signatures were different from those in the HCCs, visible mainly in the shift of the most frequent mutations from TGC > TTC in the experimental systems to GGC > GTC in the HCCs (Figs. 2, 3, 5). This seems unlikely to stem from fundamental biological divergence in cellular susceptibility between the experimental systems and the HCCs, because the experimental spectra are all relatively similar despite comprising two different human cell lines and in vivo tumors from mice both with and without the HBsAg transgene (Fig. 5B,C; Supplemental Table S2). We also note that the signature of AFB1 mutagenesis, inferred from 397 mutations in 6.4 kb of sequence in AFB1-exposed mice (Chawanthayatham et al. 2017), is similar to the AFB1sig$_{G>N}$ reported here (cosine similarity 0.96; Chawanthayatham and colleagues show mutation frequencies per trinucleotide; for comparison, the right hand side of Fig. 6A shows AFB1sig$_{G>N}$ using the same conventions). One potential reason for differences between experimental AFB1 signatures and the computationally extracted signatures from the human HCCs might be that human exposure does not consist of pure AFB1 but rather a mixture of aflatoxins, including aflatoxins G1, B2, or G2, possibly co-occurring with other mycotoxins. Another potential reason might be differences in dosage, as HCCs might reflect lower, chronic exposures over decades rather than the short periods of high exposure in the experimental systems. A final potential reason might be interaction of aflatoxin with hepatitis; although the mice with the HBsAg transgene had spectra similar to the other experimental AFB1 spectra, this model may not recapitulate all the interactions between AFB1 and hepatitis that contribute to human hepatocellular carcinogenicity.

Beyond the differences between the experimental AFB1 spectra and the HCCs, there was considerable diversity among the mouse tumors, and there were subtle but apparently consistent differences between the spectra in HepaRG and HepG2 cells. Diversity among the mouse tumors is partly due to additional mutagenic processes that arose before and during cancer development. This is especially evident in the action of COSMIC Signature 17, probably later in tumor development, in mice M1, M2, and M6 (Supplemental Figs. S6, S8, S12). In addition, the numbers of mutations in the tumors in the AFB1-exposed mice with the HBsAg transgene were substantially lower than those in the wild-type mice, even though AFB1 exposure and age at tumor sequencing were the same in both groups. Possibly AFB1 was unevenly distributed throughout the liver, leading to varying levels of exposure and mutagenesis among cells. Under this hypothesis, we speculate that, in the HBsAg mice, oncogenesis required few AFB1 mutations in addition to the HBsAg transgene, while for wild-type mice, oncogenesis occurred only in the most heavily mutated cells. This scenario would be consistent with the observation of smaller and

far fewer tumors in AFB1-treated wild-type mice than in AFB1-treated HBsAg mice (Teoh et al. 2015).

Differences between HepaRG and HepG2 may reflect variation in the metabolism of AFB1 or in intensity of mutagenesis. HepaRG and HepG2 differ in levels of and inducibility of some of the cytochrome P450s that metabolize AFB1 to both the adduct-forming AFB1 exo-8,9-epoxide and to other, less mutagenic metabolites (Hart et al. 2010; Gerets et al. 2012). For example, levels of CYP3A4 and CYP3A5 are much higher in HepaRG than in HepG2. However, the proximal adduct-forming compound is thought to be AFB1 exo-8,9-epoxide, regardless of which cytochrome P450s are expressed. Consequently, to the best of the field's understanding, differences in P450 profiles would affect mainly the dosage of AFB1 exo-8,9-epoxide. In addition, HepaRG cultures were treated with higher concentrations of AFB1 than the HepG2 cultures. Thus, dosage effects might account for the subtle differences between HepaRG and HepG2 mutational spectra. However, we cannot exclude the possibility of other, uncharacterized variation in AFB1 mutagenesis between the two cell lines that could account for these differences.

In conclusion, our multisystem approach to linking mutational signature to carcinogen exposure generally confirmed the presence of an aflatoxin signature but also pointed to some of the complexity of aflatoxin and AFB1 carcinogenesis. Our findings further indicate that, even for a relatively mutagenic compound such as AFB1, WGS yields more robust views of mutational signatures than WES analysis, because of the higher relative stochastic variability inherent in the smaller numbers of exonic mutations (Supplemental Table S12). We propose that the described multisystem approach to the experimental study of carcinogen exposure should be developed further and can be used more broadly to understand the mechanisms of environmental mutagenesis and to provide support for molecular epidemiology studies ultimately aimed at cancer prevention.

Methods

Sources of newly sequenced human tumor and normal pairs

We studied archived paraffin blocks of six HCCs and corresponding nontumor tissues from patients previously recruited in Qidong County, China, with approval by the IARC Ethics Committee (project ref. no. 15–20) (Supplemental Table S13). All six HCCs harbored the TP53 R249S somatic mutation, and the patients were positive for HBsAg and negative for hepatitis C infection (Szymanska et al. 2009).

Expression data for analysis of transcription-strand bias

We obtained expression data as follows: HepaRG (average of accessions GSM1139508, GSM1139509, GSM1139510 from Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo/) (Ambolet-Camoit et al. 2015); HepG2 (Epigenomic Roadmap http://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz); normal human liver (Epigenomic Roadmap); and normal mouse liver (http://chromosome.sdsc.edu/mouse/download/19-tissues-expr.zip) (Shen et al. 2012).

Analysis of replication timing and replication-strand bias

For analysis of human cell lines and primary tumors, we obtained processed replication timing (Repli-seq) data that provided early and late replication regions for HepG2 from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/, accession GSM923446).

We determined replication strand based on the local maxima and minima of wavelet-smoothed signal data for HepG2 (Liu et al. 2015); we took the peaks as replication initiation zones and the valleys as replication termination zones. We inferred replication direction as proceeding from initiation to termination zones.

In silico decomposition of aflatoxin-related mutational signatures and assignment of mutational-signature activities to HCCs

To discover AFB1-related mutational signatures in experimental data and somatic mutation data from human HCCs selected for likely aflatoxin exposure, we used the nsNMF method from the R NMF package (https://cran.r-project.org/web/packages/NMF/index.html) (R Core Team 2017) with the sparsity parameter theta = 0 and nrun = 200. We studied G > N signatures in trinucleotide context in the experimental AFB1 data and in HCCs likely to have been exposed to aflatoxin. Mouse spectra were normalized to the trinucleotide frequencies of the human genome. In addition, because of the possible presence of a signature similar to COSMIC Signature 5 in the low-VAF somatic mutations in some of the mouse tumors (Supplemental Figs. S4–S9), for the AFB1-only mice, we restricted this analysis to mutations with VAFs > 0.2 to enrich for mutations due only to AFB1. We studied G > N signatures (rather than all 96 mutations in trinucleotide context) to avoid interference from other mutational processes that affect A > N mutations, including COSMIC Signatures 12, 16, and 17 (http://cancer.sanger.ac.uk/cosmic/signatures).

To assess evidence of AFB1 exposure in human HCCs, we developed a customized method based on the Lee and Seung NMF algorithm (Lee and Seung 1999), in which we fixed the W (signature) matrix and only updated the H (exposure) matrix, while incorporating the smoothing matrix from Pascual-Montano et al. (2006) (Supplemental_Code). We applied this to human HCCs with the set of signatures previously found in HCCs (COSMIC Signatures 1, 4, 5, 6, 12, 16, 17, 22, 23) (http://cancer.sanger.ac.uk/cosmic/signatures), plus the two AFB1-related signatures identified in this study, with sparsity parameter theta = 0.

P values for generation of a mutational spectrum from a signature

We calculated P values for the observation of a particular observed mutational spectrum, consisting of n mutations, under the null hypothesis that it was generated from a particular mutational signature as follows. We used the rMultinom function in R to create 10,000 synthetic spectra of n mutations each, drawn with the probabilities specified by the mutational signature (R Core Team 2017). For each replicate, we calculated the cosine similarity between the synthetic spectrum and the signature. We took as the P value the proportion of replicates in which the cosine similarity was less than the cosine similarity between the signature and the observed spectrum.

Supplemental Methods describes other standard materials and protocols, including human cell lines, mouse models, whole-genome and -exome sequencing and variant calling, principal components analysis, and analysis of variant allele frequencies.

Data access

WGS reads for the human cell lines and mouse tumors from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under accession number PRJEB20736. WES reads for HCCs and nontumor samples from Qidong from this study have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) under accession number EGAS00001002490.

Acknowledgments

References

Alexandrov LB, Stratton MR. 2014. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24:** 52–60.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500:** 415–421.

Ambolet-Camoit A, Ottolenghi C, Leblanc A, Kim MJ, Letourneur F, Jacques S, Cagnard N, Guguen-Guillouzo C, Barouki R, Aggerbeck M. 2015. Two persistent organic pollutants which act through different xenosensors (α-endosulfan and 2, 3, 7, 8 tetrachlorodibenzo-p-dioxin) interact in a mixture and downregulate multiple genes involved in human hepatocyte lipid and glucose. *Biochimie* **116:** 79–91.

Bailey EA, Iyer RS, Stone MP, Harris TM, Essigmann JM. 1996. Mutational properties of the primary aflatoxin B1-DNA adduct. *Proc Natl Acad Sci* **93:** 1535–1539.

Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, et al. 2016. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538:** 260–264.

Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, Slocum SL, Wogan GN, Croy RG, Essigmann JM. 2017. Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci* **114:** E3101–E3109.

Chisari F, Pinkert C, Milich D, Filippi P, McLachlan A, Palmiter R, Brinster R. 1985. A transgenic mouse model of the chronic hepatitis B surface antigen carrier state. *Science* **230:** 1157–1160.

Chittmittrapap S, Chieochansin T, Chaiteerakij R, Treeprasertsuk S, Klaikaew N, Tangkijvanich P, Komolmit P, Poovorawan Y. 2013. Prevalence of aflatoxin induced p53 mutation at codon 249 (R249s) in hepatocellular carcinoma patients with and without hepatitis B surface antigen (HBsAg). *Asian Pac J Cancer Prev* **14:** 7675–7679.

Conaway JW, Conaway RC. 1999. Transcription elongation and human disease. *Annu Rev Biochem* **68:** 301–319.

Egner PA, Groopman JD, Wang J-S, Kensler TW, Friesen MD. 2006. Quantification of aflatoxin-B$_1$-N^7-guanine in human urine by high-performance liquid chromatography and isotope dilution tandem mass spectrometry. *Chem Res Toxicol* **19:** 1191–1195.

Foster P, Eisenstadt E, Miller J. 1983. Base substitution mutations induced by metabolically activated aflatoxin B1. *Proc Natl Acad Sci* **80:** 2695–2698.

Fousteri M, Mullenders LH. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res* **18:** 73–84.

Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, et al. 2016. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* **48:** 500–509.

Gerets HH, Tilmant K, Gerin B, Chanteux H, Depelchin BO, Dhalluin S, Atienzar FA. 2012. Characterization of primary human hepatocytes, HepG2 cells, and HepaRG cells at the mRNA level and CYP activity in response to inducers and their predictivity for the detection of human hepatotoxins. *Cell Biol Toxicol* **28:** 69–87.

Groopman JD, Donahue PR, Zhu J, Chen J, Wogan GN. 1985. Aflatoxin metabolism in humans: detection of metabolites and nucleic acid adducts in urine by affinity chromatography. *Proc Natl Acad Sci* **82:** 6492–6496.

Guengerich FP, Johnson WW, Shimada T, Ueng Y-F, Yamazaki H, Langouët S. 1998. Activation and detoxication of aflatoxin B 1. *Mutat Res* **402:** 121–128.

Hart SN, Li Y, Nakamoto K, Subileau EA, Steen D, Zhong XB. 2010. A comparison of whole genome gene expression profiles of HepaRG cells and HepG2 cells to primary human hepatocytes and human liver tissues. *Drug Metab Dispos* **38:** 988–994.

Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15:** 585–598.

Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R, et al. 2013. Mutational signature of aristolochic acid as revealed by whole-exome sequencing. *Sci Transl Med* **5:** 197ra102.

Hollstein M, Sidransky D, Vogelstein B, Harris CC. 1991. p53 mutations in human cancers. *Science* **253:** 49–53.

Hollstein M, Alexandrov LB, Wild CP, Ardin M, Zavadil J. 2016. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene* **36:** 158–167.

Hu J, Adar S, Selby CP, Lieb JD, Sancar A. 2015. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev* **29:** 948–960.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. 2012. *A review of human carcinogens. Part F: Chemical agents and related occupations*. International Agency for Research on Cancer, Lyon, France.

Kensler TW, Qian GS, Chen JG, Groopman JD. 2003. Translational strategies for cancer prevention in liver. *Nat Rev Cancer* **3:** 321–329.

Kensler TW, Roebuck BD, Wogan GN, Groopman JD. 2011. Aflatoxin: a 50-year odyssey of mechanistic and translational toxicology. *Toxicological Sciences* **120:** S28–S48.

Kew MC. 2013. Synergistic interaction between aflatoxin and hepatitis B virus in hepatocarcinogenesis. In *Aflatoxins—recent advances and future prospects* (ed. Razzaghi-Abyaneh M). InTech Open Access Publisher, Rijeka, Croatia.

Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401:** 788–791.

Liu F, Ren C, Li H, Zhou P, Bo X, Shu W. 2015. De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics* **32:** 641–649.

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR, et al. 2014. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* **24:** 1624–1636.

Ming L, Thorgeirsson SS, Gail MH, Lu P, Harris CC, Wang N, Shao Y, Wu Z, Liu G, Wang X, et al. 2002. Dominant role of hepatitis B virus and cofactor role of aflatoxin in hepatocarcinogenesis in Qidong, China. *Hepatology* **36:** 1214–1220.

Montesano R, Hainaut P, Wild C. 1997. Hepatocellular carcinoma: from gene to public health. *J Natl Cancer Inst* **89:** 1844–1851.

Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH. 2015. The genome as a record of environmental exposure. *Mutagenesis* **30:** 763–770.

Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, McKay J, Nedelko T, Muehlbauer KR, Marusawa H, et al. 2014. Modelling mutational landscapes of human cancers in vitro. *Sci Rep* **4:** 4482.

Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD. 2006. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans Pattern Anal Mach Intell* **28:** 403–415.

Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, Weng WH, Siew EY, Liu Y, Heng HL, et al. 2013. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* **5:** 197ra101.

Poon SL, McPherson JR, Tan P, Teh BT, Rozen SG. 2014. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med* **6:** 24.

Poon SL, Huang MN, Yang C, McPherson JR, Yu W, Heng HL, Gan A, Myint SS, Siew EY, Ler LD, et al. 2015. Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med* **7:** 38.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, et al. 2015. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* **47:** 505–511.

Severson PL, Vrba L, Stampfer MR, Futscher BW. 2014. Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human

mammary epithelial cells. *Mutat Res Genet Toxicol Environ Mutagen* **775–776:** 48–54.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488:** 116–120.

Sun G, Wang S, Hu X, Su J, Zhang Y, Xie Y, Zhang H, Tang L, Wang J-S. 2011. Co-contamination of aflatoxin B1 and fumonisin B1 in food and human dietary exposure in three areas of China. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess* **28:** 461–470.

Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, et al. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* **44:** 765–769.

Szymanska K, Chen JG, Cui Y, Gong YY, Turner PC, Villar S, Wild CP, Parkin DM, Hainaut P. 2009. TP53 R249S mutations, exposure to aflatoxin, and occurrence of hepatocellular carcinoma in a cohort of chronic hepatitis B virus carriers from Qidong, China. *Cancer Epidemiol Biomarkers Prev* **18:** 1638–1643.

Teoh WW, Xie M, Vijayaraghavan A, Yaligar J, Tong WM, Goh LK, Sabapathy K. 2015. Molecular characterization of hepatocarcinogenesis using mouse models. *Dis Model Mech* **8:** 743–753.

Wang J, Liu XM. 2007. Contamination of aflatoxins in different kinds of foods in China. *Biomed Environ Sci* **20:** 483–487.

Wild C, Garner R, Montesano R, Tursi F. 1986. Aflatoxin B1 binding to plasma albumin and liver DNA upon chronic administration to rats. *Carcinogenesis* **7:** 853–858.

Zamborszky J, Szikriszt B, Gervai JZ, Pipek O, Poti A, Krzystanek M, Ribli D, Szalai-Gindl JM, Csabai I, Szallasi Z, et al. 2016. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36:** 746–755.

Zhang W, He H, Zang M, Wu Q, Zhao H, Lu L-l, Ma P, Zheng H, Wang N, Zhang Y. 2017. Genetic features of aflatoxin-associated hepatocellular carcinomas. *Gastroenterology* **153:** 249–262.e2.