CANCER

# Mutational Signature of Aristolochic Acid Exposure as Revealed by Whole-Exome Sequencing

Margaret L. Hoang,[1] Chung-Hsin Chen,[2] Viktoriya S. Sidorenko,[3] Jian He,[1] Kathleen G. Dickman,[3,4] Byeong Hwa Yun,[5] Masaaki Moriya,[3] Noushin Niknafs,[6] Christopher Douville,[6] Rachel Karchin,[6] Robert J. Turesky,[5] Yeong-Shiau Pu,[2] Bert Vogelstein,[1] Nickolas Papadopoulos,[1] Arthur P. Grollman,[3,4] Kenneth W. Kinzler,[1]* Thomas A. Rosenquist[3]*

**In humans, exposure to aristolochic acid (AA) is associated with urothelial carcinoma of the upper urinary tract (UTUC). Exome sequencing of UTUCs from 19 individuals with documented exposure to AA revealed a remarkably large number of somatic mutations and an unusual mutational signature attributable to AA. Most of the mutations (72%) in these tumors were A:T-to-T:A transversions, located predominantly on the nontranscribed strand, with a strong preference for deoxyadenosine in a consensus sequence (T/CAG). This trinucleotide motif overlaps the canonical splice acceptor site, possibly accounting for the excess of splice site mutations observed in these tumors. The AA mutational fingerprint was found frequently in oncogenes and tumor suppressor genes in AA-associated UTUC. The AA mutational signature was observed in one patient's tumor from a UTUC cohort without previous indication of AA exposure. Together, these results directly link an established environmental mutagen to cancer through genome-wide sequencing and highlight its power to reveal individual exposure to carcinogens.**

## INTRODUCTION

Aristolochic acid (AA) is a nitrophenanthrene carboxylic acid found in all members of the genus *Aristolochia* (*1*), plants that have been used for medicinal purposes for more than 2000 years (*2*). Remarkably, the profound nephrotoxicity and carcinogenicity associated with the use of these plants came to light only recently (*3*), when *Aristolochia fangchi*, administered to 1800 otherwise healthy Belgian women as part of a weight reduction regimen, resulted in more than 100 cases of chronic tubulointerstitial disease progressing to end-stage renal failure (*4*). Over the next several years, many of the affected women developed neoplastic changes in the upper urinary tract (*5*, *6*). The presence of aristolactam (AL)–DNA adducts, products of AA metabolism, in the kidneys and ureters in these individuals (*6*) provides a tangible link between human exposure to AA and its carcinogenic effects.

In addition to the toxicities of AA associated with the use of herbal medicines, AA has been implicated as an environmental carcinogen. Balkan endemic nephropathy (BEN), a devastating kidney disease associated with urothelial carcinoma of the upper urinary tract (UTUC), occurs exclusively in residents of rural communities in the Danube River basin (*7*). The etiology of BEN remained a mystery until AA was shown to be the causative agent of this disease (*8*–*10*). In these studies, AL-DNA adducts and the *TP53* mutational spectrum of UTUC (see below) were used as biomarkers of internal exposure to AA. Environmental exposure to AA was attributed to consumption of wheat grain contaminated with seeds of *Aristolochia clematitis* (*11*, *12*).

Epidemiologic studies in Taiwan, using a national prescription database, reveal that, between 1997 and 2003, about one-third of the Taiwanese population had been prescribed remedies containing *Aristolochia* herbs (*13*). Moreover, the incidence of UTUC in Taiwan is the highest in the world (*14*). As in the Balkans, the detection of AL-DNA adducts and the spectrum of *TP53* mutations in the associated urothelial cancers supported the designation of AA as a major cause of UTUC in Taiwan (*15*).

The mutational spectrum of *TP53* in UTUC associated with AA exposure in both the Balkans and Taiwan is dominated by A:T-to-T:A mutations (*8*, *9*, *15*, *16*). Likewise, the predominant (48%) mutations seen in the human *TP53* gene in knock-in mouse cells treated in vitro with AA are A:T-to-T:A transversions (*17*). In contrast, among the 27,000 mutations in the International Agency for Research on Cancer (IARC) *TP53* database, A:T-to-T:A *TP53* mutations are found in 5.3% of all human cancers and only 1.4% of UTUCs overall (*18*). The altered spectrum of *TP53* mutations associated with exposure to AA suggests that AA is acting as the causative agent. Supporting this idea, translesional DNA synthesis past AL-deoxyadenosine-DNA adducts preferentially fosters misincorporation of deoxyadenosine, leading to A-to-T transversions (*19*). The demonstration that AL-DNA adducts are strongly resistant to global genomic nucleotide excision repair (*20*) accounts for the strand bias seen in *TP53* mutations and contributes to their remarkable persistence in human tissues (*6*, *9*).

The studies reviewed above provide important insights into AA-associated cancers but leave a number of questions unanswered. For example, does AA exposure result in more mutations per tumor or just an altered spectrum of mutations? Are the *TP53* mutations associated with AA representative of mutations throughout the genome? And what mutated genes beyond *TP53* drive AA-associated UTUC? To help answer these questions, we characterized the somatic muta-

[1]Ludwig Center for Cancer Genetics and Therapeutics and the Howard Hughes Medical Institute at Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA. [2]Department of Urology, National Taiwan University Hospital and College of Medicine, Taipei 10002, Taiwan. [3]Department of Pharmacological Sciences, Stony Brook University, Stony Brook, NY 11794, USA. [4]Department of Medicine, Stony Brook University, Stony Brook, NY 11794, USA. [5]Division of Environmental Health Sciences, Wadsworth Center, New York State Department of Health, Empire State Plaza, Albany, NY 12201, USA. [6]Department of Biomedical Engineering, Institute for Computational Medicine, the Johns Hopkins University, Baltimore, MD 21218, USA. *Corresponding author. E-mail: kinzlke@jhmi.edu (K.W.K.); thomas.rosenquist@stonybrook.edu (T.A.R.)

tions in all of the protein-coding genes in UTUCs from 26 individuals, including 19 with suspected exposure to AA.

## RESULTS

### Exome sequencing of UTUCs

The exomes of 19 UTUCs and matched normal tissue from individuals with documented AA exposure were sequenced. Biomarkers used to indicate patient exposure to AA were the presence of AL-DNA adducts in kidney DNA, as determined by the $^{32}$P-postlabeling method, and/or a rare A:T-to-T:A mutation in tumor *TP53* gene (*15*, *21*) (table S1). For comparison, we sequenced 7 UTUCs from individuals with no suspected AA exposure; all of these individuals were smokers, whereas all 19 AA-UTUCs were never-smokers. Both the AA-UTUC and the smoking-associated (SA)–UTUC cohorts were selected from the Taiwanese population and were similar in age (average, 69 versus 64) and extent of disease (Table 1 and table S1). The average coverage of each base in the targeted regions was high (130- and 111-fold for the AA and SA cohorts, respectively), and 89% of the 37,907,452 bases of coding exons were covered by at least 10 reads (Table 1 and table S2).

Analysis of these data revealed an average of 17,102 known single-nucleotide polymorphisms (SNPs) per individual. These SNPs allowed two important quality controls. First, the tumor and normal tissue from each individual shared, on average, 99.96% of the SNPs, confirming their origin from the same patient. Second, the SNPs allowed identification of regions containing allelic imbalance because of either loss of heterozygosity (LOH) or gain of a chromosomal segment. Using regions of LOH, we were able to confirm a high neoplastic cell content (>45%) in 24 of 26 samples (table S1). The remaining two samples were subsequently confirmed to have a high neoplastic cell content (>60%) on the basis of the allelic ratios of somatic mutations (table S1).

### Identification of somatic mutations

Using stringent criteria to avoid false-positive calls (see Materials and Methods), we identified a total of 14,957 mutations in the two cohorts, with a median of 304 (range, 87 to 3303) somatic mutations in AA-UTUCs versus 92 (range, 20 to 178) in SA-UTUCs (tables S1 and S3). The vast majority (98.4 and 97.8% in the AA and SA cohorts, respectively) of these mutations were single-base substitutions (SBSs). The accuracy of the somatic mutation calls was independently validated by Sanger sequencing of a subset of somatic mutations. Two hundred eighty-six of the 297 (96%) tested somatic mutations were confirmed by this analysis (tables S3 and S4). There was a median of 233 (range, 69 to 2573) nonsynonymous mutations in AA-UTUCs versus 70 (range, 19 to 143) in SA-UTUCs (see table S1). The high mutation load observed in AA-UTUCs (mean, 753 mutations per tumor) is consistent

**Table 1. Summary of exome sequence analysis of human urothelial carcinomas of the upper urinary tract.**

| | | UTUC cohort | |
|---|---|---|---|
| | | AA-associated | Smoking-associated |
| Characteristics of sample analyzed | Male | 11 | 6* |
| | Female | 8 | 0 |
| | Average age at diagnosis (years) | 69 | 64 |
| Sequencing | Bases sequenced (after quality filtering) | 540,388,137,550 | 198,257,916,200 |
| | Bases mapped to genome | 460,243,168,050 | 186,169,938,500 |
| | Bases mapped to targeted region | 236,240,310,243 | 85,724,475,798 |
| | Average no. of reads per targeted base | 130 | 111 |
| | Targeted bases with at least 10 reads (%) | 89% | 89% |
| | Known SNPs identified in targeted region† | 649,783 | 239,525 |
| Somatic mutations | Total somatic mutations | 14,305 | 547 |
| | Insertions or deletions | 227 | 12 |
| | Single-base substitutions | 14,078 | 535 |
| Mutation description | Synonymous | 3,197 | 124 |
| | Missense | 9,377 | 368 |
| | Nonsense | 905 | 35 |
| | Frameshift | 225 | 12 |
| | Splice acceptor | 486 | 6 |
| | Splice donor | 68 | 1 |
| | Nonstop | 47 | 1 |
| Frequency of A:T>T:A | | 72% | 7% |

*Patient SA_116 excluded from the SA cohort because of AA exposure found during this study (see main text and Fig. 3). †dbSNP (*53*).

with exposure to a potent mutagen and considerably greater than the average number of mutations observed in ultraviolet (UV)–induced melanomas and smoking-induced lung cancers (*22*).

On average, 524 genes were mutated in each AA-associated tumor, making it difficult to distinguish mutated driver from mutated passenger genes. Nevertheless, several known driver genes (*22*) were found to be mutated in these UTUCs, including genes previously implicated in UTUCs, such as *FGFR3* (8%), *HRAS* (4%), *NRAS* (15%), and *TP53* (58%) (Table 2 and table S1). Nonsynonymous mutations were frequently found in driver genes involved in the chromatin modification pathway (fig. S1). These genes include *MLL2* (62%), *CREBBP* (38%), and *KDM6A/UTX* (15%), which have been noted previously in cancers of the bladder but not in UTUC (*23*). Finally, several driver genes not previously associated with tumors of the urinary tract were found to be frequently mutated in the UTUC cases studied here. These included *STAG2* (27%) and *BRCA2* (19%).

## The mutational signature of AA

The SBSs in the AA cohort showed a marked mutagenic signature, with A:T>T:A transversions accounting for 73% of SBSs (Fig. 1). A:T>T:A

**Table 2. Genetic characteristics of human urothelial carcinomas of the upper urinary tract.** Gene list was compiled using the following criteria: (i) commonly mutated oncogene or tumor suppressor gene, specifically genes listed in table S2 of (*22*); (ii) nonsynonymous mutation found in three or more UTUCs and/or recurrent nonsynonymous mutation observed in two or more UTUCs; and (iii) if no recurrent mutations, then at least one nonsynonymous mutation is potentially inactivating (nonsense, frameshift, splice site). Twenty of the possible 6940 genes with at least one nonsynonymous mutation in a UTUC satisfied the above criteria and are listed here. nd, not detected.

| Frequency | | Number of mutations | | | | | Recurrent mutations* | | AA mutational signature† | |
|---|---|---|---|---|---|---|---|---|---|---|
| Known driver gene | Number of tumors with nonsynonymous mutation (frequency‡) | Nonsynonymous§ | | | | Synonymous | Complementary DNA | Protein | Frequency of nonsynonymous mutations with AA signature¶ | |
| | | Missense | Nonsense | Frameshift | Splice site | | | | AA-associated UTUC‖ | SA-UTUC** |
| *MLL2* | 16 (62%) | 7 | 6 | 4 | 4 | 2 | | | 47% (7/15) | 17% (1/6) |
| *TP53* | 15†† (58%) | 8 | 3 | 1 | 6 | 0 | c.840A>T | p.R280S | 93% (13/14)‡‡ | 0% (0/4)‡‡ |
| *CREBBP* | 10 (38%) | 7 | 2 | 0 | 4 | 0 | IVS28-2A>T | N/A | 82% (9/11) | 0% (0/2) |
| *STAG2* | 7 (27%) | 4 | 1 | 2 | 2 | 1 | | | 43% (3/7) | 0% (0/2) |
| *BRCA2* | 5 (19%) | 3 | 2 | 0 | 0 | 0 | | | 100% (5/5) | nd |
| *KDM6A* | 4 (15%) | 0 | 3 | 1 | 0 | 0 | | | 0% (0/3) | 0% (0/1) |
| *ATRX* | 4 (15%) | 2 | 2 | 0 | 0 | 1 | | | 67% (2/3) | 0% (0/1) |
| *ASXL1* | 4 (15%) | 3 | 0 | 1 | 0 | 0 | | | 25% (1/4) | nd |
| *MLL3* | 4 (15%) | 3 | 0 | 1 | 0 | 0 | | | 33% (1/3) | 0% (0/1) |
| *SMARCA4* | 4 (15%) | 5 | 1 | 0 | 0 | 0 | | | 100% (6/6) | nd |
| *BCOR* | 4 (15%) | 3 | 0 | 0 | 1 | 1 | | | 50% (2/4) | nd |
| *NRAS* | 4 (15%) | 4 | 0 | 0 | 0 | 0 | c.182A>T | p.Q61L | 100% (4/4) | nd |
| *ARID1A* | 3 (12%) | 1 | 3 | 0 | 0 | 2 | | | 100% (4/4) | nd |
| *ABL1* | 3 (12%) | 2 | 1 | 0 | 0 | 1 | | | 100% (3/3) | nd |
| *ARID1B* | 3 (12%) | 2 | 1 | 0 | 0 | 1 | | | 33% (1/3) | nd |
| *ARID2* | 3 (12%) | 3 | 1 | 0 | 0 | 2 | | | 75% (3/4) | nd |
| *NCOR1* | 3 (12%) | 2 | 1 | 0 | 0 | 0 | | | 100% (3/3) | nd |
| *AKT1* | 2 (8%) | 2 | 0 | 0 | 0 | 0 | c.49G>A | p.E17K | 0% (0/1) | 0% (0/1) |
| *PIKC3A* | 2 (8%) | 2 | 0 | 0 | 0 | 0 | c.1633G>A | p.E545K | nd | 0% (0/2) |
| *FGFR3* | 2 (8%) | 2 | 0 | 0 | 0 | 0 | c.746C>G | p.S249C | 0% (0/1) | 0% (0/1) |
| Total | | 65 | 27 | 10 | 17§§ | 11 | | | 68% (67/98) | 5% (1/20) |

*Observed in two or more tumors. †A>T on nontranscribed strand. Note that A-T transversions are the most rare SBS observed in COSMIC data set (see Fig. 1). ‡Frequency calculated as number of UTUCs with nonsynonymous mutation in the driver gene divided by 26 (total UTUCs analyzed). §No nonstop mutations detected in this gene set, so nonstop subcategory omitted. ¶Frequency calculated as number of nonsynonymous A>T mutations on nontranscribed strand in UTUC cohort divided by the total number of nonsynonymous mutations in UTUC cohort (raw numbers are in parentheses). ‖Twenty UTUCs include 19 from AA cohort and 1 patient (SA_116) for whom AA association was subsequently found in this study (see Fig. 3). **Six UTUCs from SA cohort, excluding patient SA_116 (see footnote ‖). ††Nonsynonymous mutations detected in three UTUCs (AA_101, AA_106, and AA_109) by Sanger sequencing (see table S1). ‡‡Frequency may be skewed because of selection of UTUCs with and without A>T mutation in *TP53* for AA and SA cohorts (see Materials and Methods). §§Sixteen of 17 in splice acceptor and 1 of 17 in splice donor.

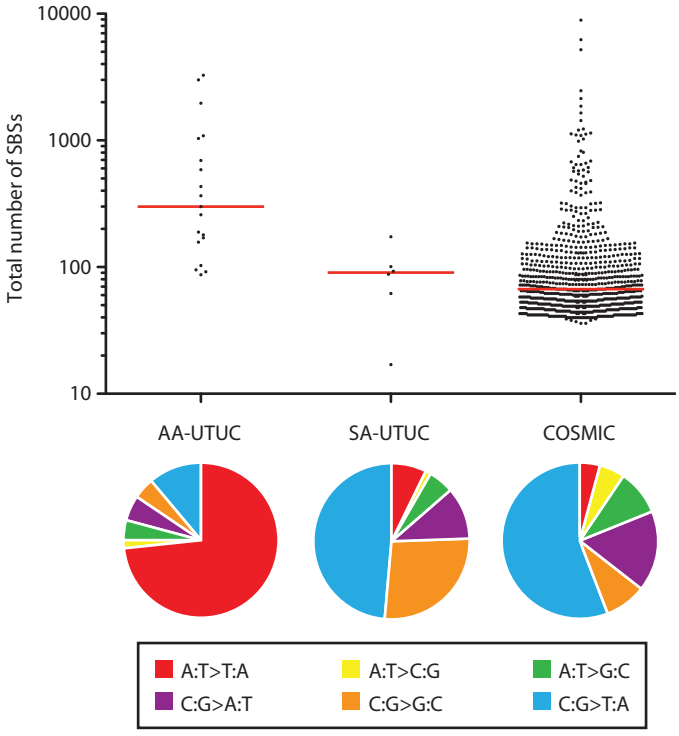


**Fig. 1. Mutation spectrum of AA-associated UTUCs.** (**Top**) Scatter plot of total number of SBSs found from exome sequencing of two UTUC cohorts: AA-UTUC and SA-UTUC. The COSMIC data set represents a broad distribution of cancer types (details in tables S5 and S6); mutations are in exomic regions for comparison with UTUC exomes. Each dot indicates an individual tumor (19 tumors for AA-UTUCs, 6 tumors for SA-UTUCs, and 812 tumors for COSMIC data set). One tumor from SA-UTUCs (SA_116) was excluded from the SA group calculations because of the presence of the AA mutational signature (details in text and Fig. 3). Only COSMIC tumors with ≥40 SBSs were included to reflect each tumor mutation spectrum. Horizontal red bars indicate median. (**Bottom**) Pie charts of SBS mutation frequencies in the exomes of AA-UTUC, SA-UTUC, and COSMIC groups. Pie legend contains the six possible SBSs on double-stranded DNA. For example, A:T>T:A is both A-to-T and T-to-A mutations, where colon (:) represents the bond between DNA strands. Group frequencies of A:T>T:A transversions (red in pie graph) are 73, 7, and 4% of SBSs in AA-UTUC, SA-UTUC, and COSMIC groups, respectively. A:T>T:A transversions are the least-frequent SBS in the COSMIC data set.

transversions are uncommon in other cancers [4.4% of 119,825 SBSs in the Catalog of Somatic Mutations in Cancer (COSMIC) database (Fig. 1 and tables S5 and S6) (*24*)] but have been observed frequently in *TP53* in AA-UTUCs (*2*). This pattern of mutations was distinct from that observed in other hypermutable tumors, including those associated with UV exposure (*25*) and tobacco smoke (*26*, *27*) and those associated with defects in *POLE* or microsatellite instability (*28*, *29*). The lack of a microsatellite instability signature in these samples is of interest given a previous report of microsatellite instability in sporadic upper urinary tract cancer (*30*).

The A:T>T:A signature within the coding regions of the AA cohort was biased more than twofold toward the nontranscribed strand (Fig. 2A). However, this strand bias is less than that previously reported for *TP53* mutations (*9*, *15*), presumably due to the inclusion of nontranscribed genes in our genome-wide data. Such genes would not be subject to transcription-coupled DNA repair. We also observed a preference for a C or T in the base preceding the mutated A residue, and a preference for G at the following base in both synonymous and nonsynonymous A>T mutations (Fig. 2B and fig. S2). This preference coincides with the canonical splice acceptor sequence of the nontranscribed strand (that is, T/CAG) and likely accounts for the 7.4-fold overrepresentation of splice acceptor mutations we observed in AA-UTUCs (Fig. 2C). It also is consistent with the previously reported overrepresentation of acceptor splice site mutations in *TP53* (*16*). Additionally, the AA mutational signature was present in oncogenes and tumor suppressor genes frequently mutated in UTUCs, where 67 of 98 (68%) nonsynonymous mutations were A>T mutations on the nontranscribed strand in AA-associated UTUCs compared to 1 of 20 (5%) nonsynonymous mutations in SA-UTUCs (Table 2).

In contrast, examination of the distribution of SBSs in the SA cohort revealed that the predominant mutations were C:G>T:A transitions (48% of SBSs, see Fig. 1). This pattern differs from the C:G>A:T transversions commonly observed in lung cancers (*26*, *27*). This paucity of C:G-to-A:T mutations suggests that tobacco carcinogens, such as polycyclic aromatic hydrocarbons including benzo[*a*]pyrene or reactive aldehydes, such as acrolein, do not play a major role in SA-UTUC, but others, such as certain nitrosamines, may be involved (*31*).

### The AA mutational signature as a marker of AA exposure

In addition to the global patterns of mutations distinguishing the AA from the SA cohorts, we examined the mutation patterns in each individual UTUC (Fig. 3A). Seventeen of the 19 AA-UTUCs had a percentage of A:T>T:A transversions that was more than six times the interquartile range (IQR) of COSMIC tumors away from the COSMIC median (median of 5.6% of SBSs with IQR of 6.2%), strongly supporting exposure to AA as the causative event in these cancers (Fig. 3B). This allowed the formulation of an AA signature defined as a mutation load of ≥40 SBS mutations and >35% A:T>T:A transversions. This signature was not observed in 812 tumors with 40 or more SBS mutations present in COSMIC (Fig. 3B). In two AA-UTUC cases, AA_101 and AA_129, the percentages of A:T>T:A transversions were not consistent with the AA signature, making the etiologic role of AA in these patients less clear (Fig. 3, A and B).

Surprisingly, this analysis revealed that one of the SA patients (SA_116) had a highly elevated level of A:T>T:A transversions [55 of 102 SBSs (54%)], consistent with the AA signature (Fig. 3B). This level is within the IQR of the AA-UTUCs (IQR between 48 and 79%) but is almost twofold greater than the highest value (29%) observed in the 812 tumors from COSMIC. Furthermore, as observed in AA-induced cancers, A:T>T:A transversions showed a bias for the nontranscribed strand (Fig. 3C) and a preference for a CAG context (Fig. 3D).

AL-DNA adducts were undetectable in kidney DNA from patient SA_116, as determined by $^{32}$P-postlabeling analysis. To investigate this apparent discrepancy, we applied a recently developed, ultrasensitive mass spectrometric method (*32*) for the determination of AL-DNA adducts to nine patients who were adduct-negative by $^{32}$P-postlabeling techniques, which includes two from the AA cohort and all seven from the SA cohort (see Materials and Methods). Consistent with the documented, widespread use of *Aristolochia*-containing herbal medicines in Taiwan (*13*, *14*), mass spectrometric analysis revealed that eight of the nine renal DNA samples had detectable AL-DNA adducts. Patient SA_116 had 5.5 adducts per $10^8$ nucleotides, a concentration similar

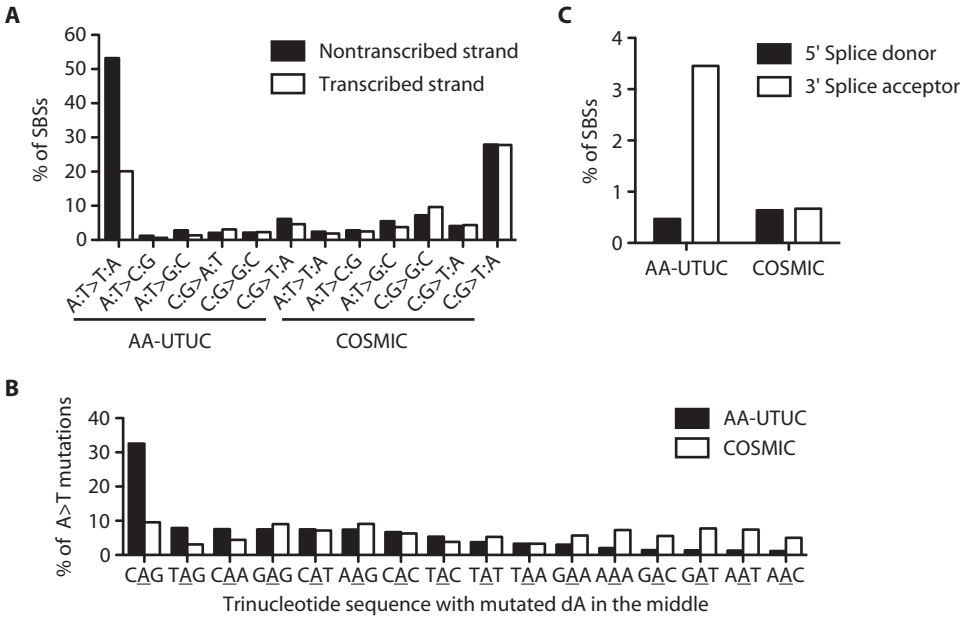


**Fig. 2. Mutational pattern of A:T>T:A transversions in AA-UTUC exomes.** (**A**) Frequencies of the six possible SBSs listed on the *x* axis found on the nontranscribed strand (black bars) or transcribed strand (white bars) in AA-UTUC (left side) and COSMIC data set (right side). The A:T>T:A strand bias in AA-UTUCs likely represents an enrichment of A>T mutations on the nontranscribed strand (rather than T>As on the transcribed strand). (**B**) Frequencies of A>T mutations within each observed trinucleotide sequence in AA-UTUCs (black bars) and COSMIC data set (white bars). The middle A is the mutated base (A>T) in trinucleotide sequences listed on the *x* axis. Total number of A>Ts evaluated is 10,326 for AA-UTUCs and 5250 for COSMIC data set. (**C**) Frequencies of mutations found in 5′ splice donor (black bars) and 3′ splice acceptor (white bars) sites in AA-UTUCs and COSMIC data set. Mutations only in the canonical 5′ splice donor (GT) and 3′ splice acceptor (CAG or TAG) sites were counted. GT and CAG/TAG splice site sequences correspond to the nontranscribed strand sequences.

to or higher than three of the patients in the AA cohort, and higher than any of the other SA patients (table S1). These adduct findings are consistent with an AA exposure–induced cancer in patient SA_116. However, AL-DNA adducts indicate exposure in kidney tissue and do not reveal the timing and extent of exposure in urothelial tissue giving rise to the cancer. Although other yet to be determined carcinogens may result in similar signatures, the presence of the AA mutational signature in the tumor, coupled with the tumor type and detectable AL-DNA adducts in normal tissue, provides strong evidence for AA as a causative agent in this case.

## DISCUSSION

Whole-exome sequencing analysis of AA- and SA-UTUCs allowed us to characterize in depth the mutational signature of AA exposure and to identify drivers in this tumor type. Driver genes mutated in a large fraction of UTUCs include previously implicated genes [*HRAS* (*33*), *TP53* (*34*), and *FGFR3* (*35*)] and genes not previously reported to be mutated in UTUCs (*MLL2*, *CREBBP*, *STAG2*, *BRCA2*, *KDM6A*, and *NRAS*). Of the newly identified genes, several that were mutated in a high fraction of UTUCs (*MLL2*, *CREBBP*, and *KDM6A*) were of particular interest because they implicate defects in chromatin modification in the development of UTUCs. One or more of these three genes were mutated in over two-thirds of UTUCs, and when less frequently mutated chromatin remodeling genes (*ARID2*, *ATRX*, *MLL3*, and *PBRM1*) are considered, this figure rises to 81% (21 of 26) of UTUCs.

It has recently been shown that half or more of the somatic mutations in cancers occur during the replication of normal stem cells before the onset of neoplasia (*36*). These background passenger mutations can make it difficult to recognize the mutational signature of a carcinogen. Nevertheless, potent mutagens leading to high levels of signature mutations can be recognized in certain tumor types. Examples so far include UV-induced melanomas that display abundant C:G-to-T:A mutations in a dipyrimidine context along with CC:GG-to-TT:AA dinucleotide substitutions (*25*, *37*–*42*). Although tobacco smoke contains more than 60 carcinogens (*43*), lung tumors of tobacco smokers have a predominance of C:G-to-A:T mutations (*26*, *27*, *44*–*48*), presumably reflecting the preferential action of one or several mutagens that leave this signature. This tobacco-related signature is not evident in all tumor types associated with smoking because it is not prominent in head and neck squamous cell carcinoma (*49*) or the SA-UTUCs studied here. Our study of AA-UTUCs reveals a genome-wide mutational signature characterized by a high mutational load with an excess of A:T-to-T:A transversions and splice acceptor mutations, as well as an enrichment of A>Ts on the nontranscribed strand with an A>T preference for a T/CAG context. All aspects of this molecular signature are similar to that reported for AA-induced UTUC in the accompanying paper (*50*).

AL-DNA adducts in normal tissues serve as a sensitive biomarker of internal exposure but alone cannot establish AA as the causative agent for neoplasia in the same patient. Exposure to AA may occur subsequent to tumor initiation or may have been coincidental. Indeed, in this study, the application of an ultrasensitive and highly specific analytical method capable of detecting less than 20 adducts per diploid cell revealed that all but one member of the AA and SA cohorts combined had been exposed to some extent previously to AA. In contrast, the observation of a strong AA mutational signature among the clonal mutations distributed throughout the genome of a tumor, and specifically in driver gene mutations, provides strong evidence of causality. As an example of this principle, the AA mutational signature allowed us to implicate AA exposure in a patient in whom smoking was believed to be the causative agent of cancer. Our observations support the idea that genome-wide sequencing can illuminate the pathogenesis of cancers in individual cases or clusters of cases suspected to be caused by exposure to environmental mutagens, providing a powerful tool for molecular epidemiology.

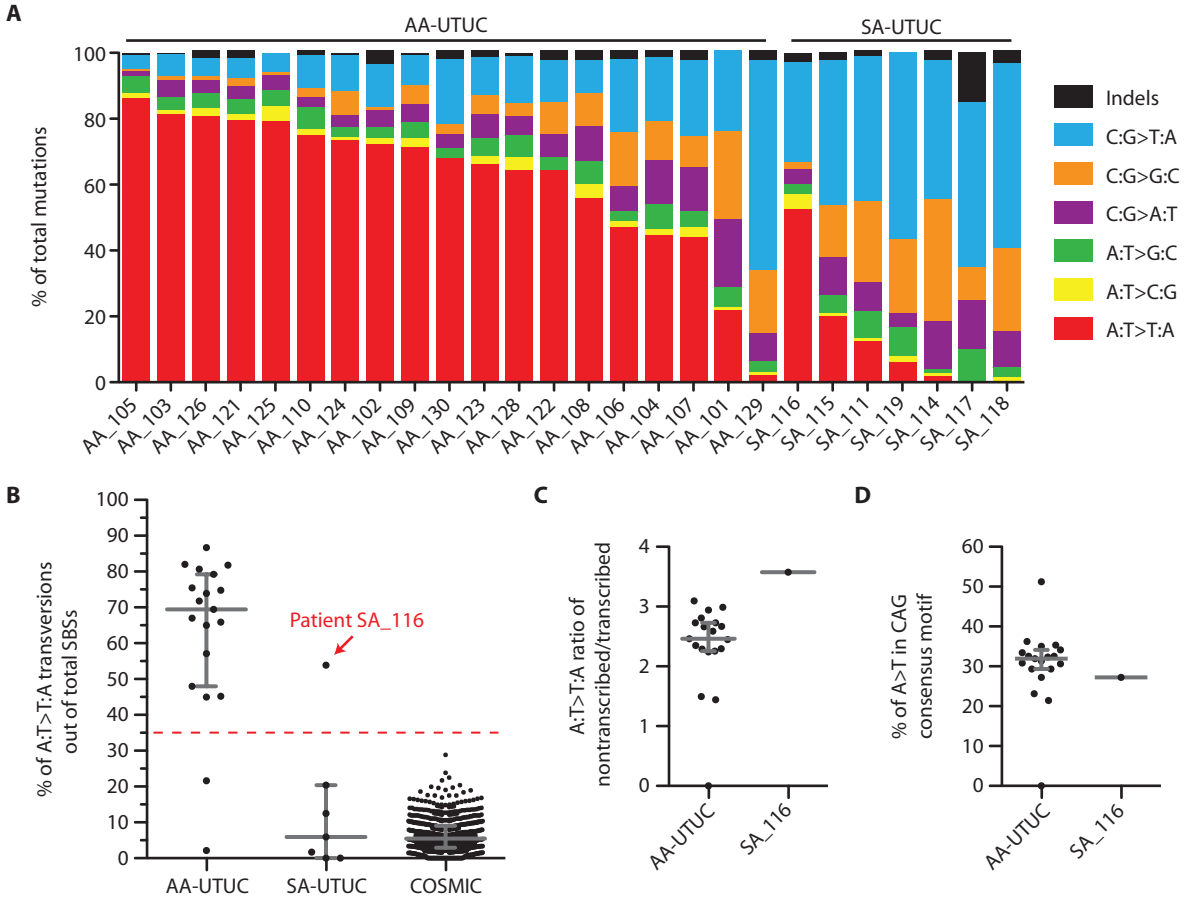


**Fig. 3. Molecular diagnosis of UTUC patient with previously unknown AA exposure.** (**A**) Mutation spectra from exome sequencing of 26 UTUC patients (labeled on the *x* axis) from AA-UTUC and SA-UTUC cohorts. AA cohort originally defined by the presence of AL-DNA adducts in kidney tissue by $^{32}$P-postlabeling method and/or the presence of A>T mutation in *TP53* (see Materials and Methods). The seven individuals in the SA cohort lacked both AA biomarkers noted above but did have a clinical history of smoking. (**B**) Dot plot of frequencies of A:T>T:A transversions out of total SBSs. Each dot is an individual tumor (19 tumors for AA-UTUC, 7 tumors for SA-UTUC, and 812 tumors for COSMIC data set). Suspected exposure to AA herein defined as above 35% (red dashed line). Median (horizontal gray lines) A:T>T:A transversion frequencies are 69% for AA-UTUC, 5.9% for SA-UTUC, and 5.6% for COSMIC data set. The IQR (indicated by vertical gray lines) is between 48 and 79% for AA-UTUC, 0 and 20% for SA-UTUC, and 2.9 and 9.1% for COSMIC data set. Patient SA_116 has 54% (55 of 102 SBSs) A:T>T:A transversions (red arrow). This 54% value lies within the IQR of AA-UTUC but is above the upper quartile (by seven times the COSMIC IQR) of the COSMIC data set. (**C**) Dot plot of the ratio of A:T>T:A mutations on nontranscribed strand over transcribed strand in patient SA_116 (ratio of 3.6) relative to AA-UTUC cohort (median ratio of 2.5 with IQR of 2.3 to 2.7). (**D**) Dot plot of frequencies of A>Ts within CAG consensus sequence out of total number of observed trinucleotide sequences in patient SA_116 (27%) relative to AA-UTUC cohort (median of 32% with IQR of 29 to 34%).



## MATERIALS AND METHODS

### Study design

The objective of this study was to determine the molecular signature of AA in UTUC DNA. Tumor-specific mutations were determined by comparing exome sequences of DNA from UTUC and nontumorous tissue obtained from a cohort of 20 patients for whom AA exposure was suspected. This mutation profile was compared to that in the COSMIC database and to UTUCs from a cohort of 10 ethnically and age-matched smokers. Inclusion in the AA cohort required the detection of AL-DNA adducts in kidney DNA by the $^{32}$P-postlabeling method and/or the presence of an A:T-to-T:A mutation in the *TP53* gene. Patients selected for the SA cohort were negative for both these parameters and had a history of smoking. After determining the mutation profiles in each cohort, we reanalyzed the AL-DNA adduct content of all kidney DNAs using a recently developed, sensitive mass spectrometric method.

### Preparation of clinical samples

The research protocol was reviewed and approved by the Institutional Review Boards of Stony Brook University, the National Taiwan University, Wadsworth Center, and Johns Hopkins University. A complete description of the clinical parameters of a cohort of 151 UTUC and kidney tissue pairs was reported previously (*15*, *21*), together with concentrations of AL-DNA adducts in the kidney DNA as determined by the $^{32}$P-postlabeling method, and mutations in the *TP53* gene in UTUC DNA (*15*, *21*). For the present study, we selected a cohort of 20 tumor patients (AA_101 to AA_110 and AA_121 to AA_130) with documented AA exposure, as defined by the presence of either AL-DNA adducts or an A>T mutation in *TP53*. Specifically, 10 of the

tumors selected had both AL-DNA adducts and an A>T mutation in *TP53*, 8 tumors had AL-DNA adducts without an A>T mutation in *TP53*, and 2 tumors had an A>T mutation in *TP53* without adducts detectable by $^{32}$P-postlabeling (table S1). Additionally, 10 tumors (SA_111 to SA_120) were selected from patients who had neither AL-DNA adducts by $^{32}$P-postlabeling nor *TP53* A-to-T mutations but did have a history of smoking. Subsequently, four tumors that were sequenced failed quality control as detailed below.

### Mass spectrometric determination of AL-DNA adducts

AL-DNA adduct concentrations were determined by ultraperformance liquid chromatography–electrospray ionization/multistage scan mass spectrometry. Five micrograms of kidney DNA from nine patients (AA_103, AA_104, SA_111, SA_114, SA_115, SA_116, SA_117, SA_118, and SA_119) that were adduct-negative by $^{32}$P-postlabeling and one control (AA_129) that was adduct-positive by $^{32}$P-postlabeling was hydrolyzed with a cocktail of enzymes containing deoxyribonuclease (DNase) I, nuclease P1, alkaline phosphatase, and phosphodiesterase, as previously described (*51*). Adduct quantification used the stable isotope dilution method, with [$^{15}N_5$]-dA-AL-I as the internal standard at a concentration of two to five adducts per $10^8$ DNA bases (*32*). Analyses were performed with a nanoACQUITY UPLC system (Waters Corporation) interfaced with an Advance CaptiveSpray source from Michrom Bioresources Inc. and an ion trap mass spectrometer (LTQ Velos, Thermo Fisher). A Waters Symmetry trap column (180 μm × 20 mm, 5-μm particle size) was used for online solid-phase enrichment of the DNA adducts. A C18 AQ (0.3 × 150 mm, 3-μm particle size, Michrom Bioresources Inc.) was used for chromatography. The DNA adducts were measured in the positive ionization mode at the $MS^3$ scan stage. The chromatographic and mass spectra acquisition parameters were previously described (*32*).

### Preparation of Illumina genomic DNA libraries

Genomic DNA libraries for cases AA_101 to AA_110 were prepared following a modified Illumina protocol as described (*52*). Libraries for the remaining cases were prepared with a TruSeq library kit (Illumina) with the following modifications. (i) Genomic DNA (0.1 to 2 μg) from tumor or normal cells in 55 μl of tris-EDTA (TE) buffer was sonicated targeting 250 base pairs (bp) (range, 100 to 500 bp) with a Bioruptor (Diagenode) at intensity H with seven cycles of 30 s on and 90 s off in 3°C water bath. (ii) Fragmented DNA (50 μl) was mixed with 40 μl of End Repair Mix and 10 μl of resuspension buffer. The 100-μl end-repair mixture was incubated at 30°C for 30 min. (iii) End-repair mixture was size-selected for ~200-bp inserts with AMPure XP beads (Agencourt). Specifically, 80 μl of AMPure XP beads was mixed with the 100-μl end-repair reaction, incubated at room temperature for 15 min, and then placed on a magnet for 5 min. The supernatant (~175 μl) was transferred to a new tube. AMPure XP beads (50 μl) were mixed with the ~175 μl of supernatant, incubated for 15 min, and placed on a magnet for 5 min. Supernatant was removed and discarded, and the beads were washed twice with 200 μl of freshly prepared 80% ethanol with 30-s room temperature incubation for each wash. Beads were left at room temperature for 15 min to dry. Beads were resuspended in 19 μl of resuspension buffer, incubated at room temperature for 2 min, and then placed on a magnetic stand for 2 min. Clear supernatant (15 μl) was transferred to a new tube. (iv) To A-tail, 15 μl of the supernatant (purified end-repaired DNA) was mixed with 2.5 μl of resuspension buffer and 12.5 μl of A-tailing mix, and incubated at 37°C for 30 min. (v) For adaptor ligation, the 30 μl of A-tailed DNA was mixed with 2.5 μl of resuspension buffer, 2.5 μl of ligation mix, and 2.5 μl of adaptor index. The ligation mixture was incubated for exactly 10 min at 30°C. After the 10-min ligation, 5 μl of stop ligation buffer was added. (vi) The adaptor-ligated DNA was purified with two rounds of AMPure XP beads with a 1:1 ratio of beads to sample. Specifically, 42.5 μl of AMPure XP beads was mixed with 42.5 μl of the ligation reaction, incubated for 15 min, and placed on magnet for 5 min. Supernatant was discarded, and beads were washed twice with freshly prepared 80% ethanol with 30-s room temperature incubation for each wash. Beads were left at room temperature for 15 min to dry. Dry beads were resuspended in 52.5 μl of resuspension buffer, incubated at room temperature for 2 min, and then placed on magnetic stand for 2 min. Clear supernatant (50 μl) was transferred to a new tube. AMPure XP beads (50 μl) were mixed with the 50-μl supernatant, and the purification scheme detailed above was repeated, except the beads were resuspended with 25 μl of resuspension buffer, and 21 μl of clear supernatant (purified adapter-ligated DNA) was then transferred to a new tube. (vii) To obtain an amplified library, 20 μl of adapter-ligated DNA was mixed with 5 μl of polymerase chain reaction (PCR) primer Cocktail and 25 μl of PCR Master Mix. PCR program used was as follows: 10 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 30 s, and then 1 cycle of 72°C for 5 min. To purify the PCR product, 48 μl of AMPure XP beads was mixed with 48 μl of PCR, incubated at room temperature for 5 min, and placed on a magnet for 5 min. Supernatant was discarded, and beads were washed twice with freshly prepared 80% ethanol with 30-s room temperature incubation for each wash. Beads were left at room temperature for 15 min to dry. Dry beads were resuspended in 12 μl of resuspension buffer, incubated at room temperature for 2 min, and placed on a magnetic stand for 2 min. Clear supernatant (10 μl) (amplified TruSeq library) was transferred to a new tube. DNA concentration of TruSeq libraries was measured with BioAnalyzer (Agilent).

### Exome and targeted subgenomic DNA capture

Human exome capture was performed following a protocol from Agilent's SureSelectXT Human All Exon V4 (catalog no. 5190-4634, Agilent) with the following modifications. (i) A hybridization mixture was prepared containing 25 μl of SureSelect Hyb #1, 1 μl of SureSelect Hyb #2, 10 μl of SureSelect Hyb #3, and 13 μl of SureSelect Hyb #4. (ii) PE-library DNA (3.4 μl, 0.5 μg) described above, SureSelect Block #1 (2.5 μl), SureSelect Block #2 (2.5 μl), and Block #3 (0.6 μl) were loaded into one well in a 384-well Diamond PCR plate (catalog no. AB-1111, Thermo Scientific), sealed with MicroAmp clear adhesive film (catalog no. 4306311; ABI), placed in GeneAmp PCR System 9700 thermocycler (Life Sciences Inc.) for 5 min at 95°C, and then held at 65°C (with the heated lid on). (iii) Hybridization buffer (25 to 30 μl) from step (i) was heated for at least 5 min at 65°C in another sealed plate with heated lid on. (iv) SureSelect Oligo Capture Library (5 μl), nuclease-free water (1 μl), and diluted RNase Block (1 μl) (prepared by diluting RNase Block 1:1 with nuclease-free water) were mixed and heated at 65°C for 2 min in another sealed 384-well plate. (v) While keeping all reactions at 65°C, 13 μl of hybridization buffer from step (iii) was added to the 7 μl of the SureSelect Capture Library Mix from step (iv) and then the entire contents (9 μl) of the library from step (ii). The mixture was slowly pipetted up and down 8 to 10 times. (vi) The 384-well plate was sealed tightly, and the hybridization mixture was incubated for 24 hours at 65°C with a heated lid.

After hybridization, five steps were performed to recover and amplify captured DNA library. (i) Magnetic beads for recovering captured DNA: 50 μl of Dynal MyOne Streptavidin C1 magnetic beads (catalog

no. 650.02, Invitrogen Dynal AS) was placed in a 1.5-ml microfuge tube and vigorously resuspended on a vortex mixer. Beads were washed three times by adding 200 μl of SureSelect binding buffer, mixing on a vortex for 5 s and then removing the supernatant after placing the tubes in a Dynal magnetic separator. After the third wash, beads were resuspended in 200 μl of SureSelect binding buffer. (ii) To bind captured DNA, the entire hybridization mixture described above (29 μl) was transferred directly from the thermocycler to the bead solution and mixed gently; the hybridization mix/bead solution was incubated in an Eppendorf thermomixer at 850 rpm for 30 min at room temperature. (iii) To wash the beads, the supernatant was removed from beads after applying a Dynal magnetic separator, and the beads were resuspended in 500 μl of SureSelect Wash Buffer #1 by mixing on a vortex mixer for 5 s and incubated for 15 min at room temperature. Wash Buffer #1 was then removed from beads after magnetic separation. The beads were further washed three times, each with 500 μl of prewarmed SureSelect Wash Buffer #2 after incubation at 65°C for 10 min. After the final wash, SureSelect Wash Buffer #2 was completely removed. (iv) To elute captured DNA, the beads were suspended in 50 μl of SureSelect elution buffer, vortex-mixed, and incubated for 10 min at room temperature. The supernatant was removed after magnetic separation, collected in a new 1.5-ml microcentrifuge tube, and mixed with 50 μl of SureSelect neutralization buffer. DNA was purified with a Qiagen MinElute column and eluted in 17 μl of 70°C elution buffer to obtain 15 μl of captured DNA library. (v) The captured DNA library was amplified in the following way: 15 PCRs, each containing 9.5 μl of $H_2O$, 3 μl of 5× Phusion HF buffer, 0.3 μl of 10 mM deoxynucleotide triphosphate, 0.75 μl of dimethyl sulfoxide, 0.15 μl of Illumina PE primer #1, 0.15 μl of Illumina PE primer #2, 0.15 μl of HotStart Phusion polymerase, and 1 μl of captured exome library, were set up. The PCR program used was as follows: 98°C for 30 s; 14 cycles of 98°C for 10 s, 65°C for 30 s, 72°C for 30 s; and 72°C for 5 min. To purify PCR products, 225 μl of PCR mixture (from 15 PCRs) was mixed with 450 μl of NT buffer from NucleoSpin Extract II kit and purified as described above. The final library DNA was eluted with 30 μl of 70°C elution buffer, and DNA concentration was estimated with BioAnalyzer (Agilent).

### Quality control of sequencing and tumor samples

The quality of sequencing and tumor samples was based on coverage and neoplastic content, respectively. For coverage, the average reads per targeted base for the 60 sequenced samples (30 tumors and 30 matched normal tissues) was 121 ± 35 (SD). Two SDs below this average were used as the cutoff for proper sequence coverage. One sample (SA_120), with an average of 16 reads per targeted base in the tumor, failed to meet this requirement and was dropped from further analysis. For neoplastic content, all heterozygous positions in the matched normal tissue were first evaluated in the tumor sample to identify regions of LOH. Theoretically, if both alleles of the SNP remained heterozygous in the tumor, the allele fraction would be 0.5. Chromosomal segments where the allele fraction "shifted" lower than ~0.5 were designated as regions of LOH, and the allele detected in these LOH regions was called the "minor allele" or the allele absent in the tumor. The percentage of neoplastic cells was estimated on the basis of the minor allele fraction. For example, a minor allele fraction of 0.2 in LOH regions represents the minor allele from nontumor cells. This implies a major allele fraction of 0.2 from nontumor cells, leaving 0.6 for the major allele from tumor cells. The number of tumor cells divided by the total number of cells is represented by 0.6/(0.6 + 0.2) = 0.75, or 75% neoplastic cell content. LOH was not detected in five samples. Of these five, one sample (SA_117) used the mutation fraction of a known driver mutation (*HRAS Q61K*) and one sample (AA_125) used the average mutation fraction of its 69 nonsynonymous mutations to estimate neoplastic content. The remaining three samples that lacked LOH (SA_112, SA_113, and SA_127) were dropped because of an insufficient number of somatic mutations (less than five mutations each) to estimate neoplastic cell content.

### Somatic mutation identification by massively parallel sequencing

Captured DNA libraries were sequenced with the Illumina GAIIx Genome Analyzer, yielding 200 (2 × 100) bp from the final library fragments. Sequencing reads were analyzed and aligned to human genome hg18 with the Eland algorithm in CASAVA 1.7 software (Illumina). A mismatched base was identified as a mutation only when (i) it was identified by more than five distinct tags with at least one read in each direction; (ii) the number of tags containing a particular mismatched base was at least 15% of the total tags; (iii) there were at least 10 distinct reads in the matched normal samples; and (iv) it was not present in >1.0% of the tags in the matched normal sample. The candidate somatic mutations were further filtered to remove any known germline variants described in dbSNP (*53*), the 1000 Genomes Project (*54*), and previous exome sequencing projects including ESP6500 (*55*).

### Confirmation of somatic mutations

All somatic mutations in the driver genes listed in Table 2, table S1, and fig. S1, along with a representative sample of mutations, were independently confirmed by Sanger sequencing of the original tumor and matched normal (tables S3 and S4). PCR amplification and Sanger sequencing were performed following protocols described previously (*56*) with the primers listed in table S4, and results are reported in tables S3 and S4.

### COSMIC analysis

A total of 122,148 SBS mutations from 812 tumors with 40 or more distinct somatic mutations (Nonstop extension, Substitution—coding silent, Substitution—Missense, Substitution—Nonsense, and Unknown) were extracted from COSMIC version 61 (table S5) (*24*). Of the 122,148 COSMIC mutations, 1711 from Unknown and 612 from Substitution categories were further excluded from analysis for the following reasons: (i) not SBS, (ii) not canonical splice donor (+1G, +2T) or canonical splice acceptor (−3C, −3T, −2A, −1G) mutation, or (iii) incomplete or ambiguous annotation. The remaining 119,825 COSMIC SBSs used to compare the SBS mutations in the UTUC samples are listed in table S6. We further verified that the mutational spectra of the highly mutated tumors in our COSMIC data set (the top 30 mutated tumors) were similar to the rest of the COSMIC group (the bottom 782 tumors).

### Generation of WebLogos

Eleven bases of sequence corresponding to five bases 5′ and five bases 3′ of A>T mutations from AA-UTUCs were uploaded to the WebLogo application found at weblogo.berkeley.edu/ (*57*). Sequences used to generate WebLogos are found in table S3.

## SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/5/197/197ra102/DC1

Fig. S1. Chromatin modification pathway mutations in UTUCs.

Fig. S2. T/CAG sequence preference at synonymous and nonsynonymous A>Ts in AA-UTUCs.

Table S1. Characteristics of human samples of urothelial carcinomas of the upper urinary tract.

Table S2. Detailed summary of sequence analysis of human urothelial carcinomas of the upper urinary tract.
Table S3. Somatic mutations in human urothelial carcinomas of the upper urinary tract.
Table S4. Primers used for Sanger sequencing.
Table S5. COSMIC studies considered.
Table S6. COSMIC mutation data set used in this study.

## REFERENCES AND NOTES


1. V. Kumar, Poonam, A. K. Prasad, V. S. Parmar, Naturally occurring aristolactams, aristolochic acids and dioxoaporphines and their biological activities. *Nat. Prod. Rep.* **20**, 565–583 (2003).
2. A. P. Grollman, J. Scarborough, B. Jelakovic, in *Advances in Molecular Toxicology*, J. Fishbein, Ed. (Elsevier, Amsterdam, 2009), vol. 3, pp. 211–222.
3. J.-L. Vanherweghem, C. Tielemans, D. Abramowicz, M. Depierreux, R. Vanhaelen-Fastre, M. Vanhaelen, M. Dratwa, C. Richard, D. Vandervelde, D. Verbeelen, M. Jadoul, Rapidly progressive interstitial renal fibrosis in young women: Association with slimming regimen including Chinese herbs. *Lancet* **341**, 387–391 (1993).
4. J. L. Vanherweghem, F. Debelle, M.-C. Muniz-Martinez, J. Nortier, in *Clinical Nephrotoxins*, M. E. de Broe, G. A. Porter, W. M. Bennet, G. A. Verpooten, Eds. (Kluwer, Dordrecht, 2003), pp. 588–601.
5. J. P. Cosyns, M. Jadoul, J. P. Squifflet, F. X. Wese, C. van Ypersele de Strihou, Urothelial lesions in Chinese-herb nephropathy. *Am. J. Kidney Dis.* **33**, 1011–1017 (1999).
6. J. L. Nortier, M. C. Martinez, H. H. Schmeiser, V. M. Arlt, C. A. Bieler, M. Petein, M. F. Depierreux, L. De Pauw, D. Abramowicz, P. Vereerstraeten, J. L. Vanherweghem, Urothelial carcinoma associated with the use of a Chinese herb (*Aristolochia fangchi*). *N. Engl. J. Med.* **342**, 1686–1692 (2000).
7. L. Djukanovic, Z. Radovanovic, in *Clinical Nephrotoxins*, M. E. de Broe, G. A. Porter, W. M. Bennet, G. A. Verpooten, Eds. (Kluwer, Dordrecht, 2003), pp. 588–601.
8. A. P. Grollman, S. Shibutani, M. Moriya, F. Miller, L. Wu, U. Moll, N. Suzuki, A. Fernandes, T. Rosenquist, Z. Medverec, K. Jakovina, B. Brdar, N. Slade, R. J. Turesky, A. K. Goodenough, R. Rieger, M. Vukelić, B. Jelaković, Aristolochic acid and the etiology of endemic (Balkan) nephropathy. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12129–12134 (2007).
9. B. Jelaković, S. Karanović, I. Vuković-Lela, F. Miller, K. L. Edwards, J. Nikolić, K. Tomić, N. Slade, B. Brdar, R. J. Turesky, Ž. Stipančić, D. Dittrich, A. P. Grollman, K. G. Dickman, Aristolactam-DNA adducts are a biomarker of environmental exposure to aristolochic acid. *Kidney Int.* **81**, 559–567 (2012).
10. M. E. De Broe, Chinese herbs nephropathy and Balkan endemic nephropathy: Toward a single entity, aristolochic acid nephropathy. *Kidney Int.* **81**, 513–515 (2012).
11. T. Hranjec, A. Kovac, J. Kos, W. Mao, J. J. Chen, A. P. Grollman, B. Jelaković, Endemic nephropathy: The case for chronic poisoning by aristolochia. *Croat. Med. J.* **46**, 116–125 (2005).
12. M. Ivić, Etiology of endemic nephropathy. *Lijec Vjesn* **91**, 1273–1281 (1969).
13. S. C. Hsieh, I. H. Lin, W. L. Tseng, C. H. Lee, J. D. Wang, Prescription profile of potentially aristolochic acid containing Chinese herbal products: An analysis of National Health Insurance data in Taiwan between 1997 and 2003. *Chin. Med.* **3**, 13 (2008).
14. M. N. Lai, S. M. Wang, P. C. Chen, Y. Y. Chen, J. D. Wang, Population-based case–control study of Chinese herbal products containing aristolochic acid and urinary tract cancer risk. *J. Natl. Cancer Inst.* **102**, 179–186 (2010).
15. C. H. Chen, K. G. Dickman, M. Moriya, J. Zavadil, V. S. Sidorenko, K. L. Edwards, D. V. Gnatenko, L. Wu, R. J. Turesky, X. R. Wu, Y. S. Pu, A. P. Grollman, Aristolochic acid-associated urothelial cancer in Taiwan. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8241–8246 (2012).
16. M. Moriya, N. Slade, B. Brdar, Z. Medverec, K. Tomic, B. Jelakovic, L. Wu, S. Truong, A. Fernandes, A. P. Grollman, *TP53* mutational signature for aristolochic acid: An environmental carcinogen. *Int. J. Cancer* **129**, 1532–1536 (2011).
17. T. Nedelko, V. M. Arlt, D. H. Phillips, M. Hollstein, *TP53* mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. *Int. J. Cancer* **124**, 987–990 (2009).
18. M. Hollstein, M. Moriya, A. P. Grollman, M. Olivier, Analysis of *TP53* mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. *Mutat. Res.* 10.1016/j.mrrev.2013.02.003 (2013).
19. S. Attaluri, R. R. Bonala, I. Y. Yang, M. A. Lukin, Y. Wen, A. P. Grollman, M. Moriya, C. R. Iden, F. Johnson, DNA adducts of aristolochic acid II: Total synthesis and site-specific mutagenesis studies in mammalian cells. *Nucleic Acids Res.* **38**, 339–352 (2010).
20. V. S. Sidorenko, J. E. Yeo, R. R. Bonala, F. Johnson, O. D. Schärer, A. P. Grollman, Lack of recognition by global-genome nucleotide excision repair accounts for the high mutagenicity and persistence of aristolactam-DNA adducts. *Nucleic Acids Res.* **40**, 2494–2505 (2012).
21. C. H. Chen, K. G. Dickman, C. Y. Huang, M. Moriya, C. T. Shun, H. C. Tai, K. H. Huang, S. M. Wang, Y. J. Lee, A. P. Grollman, Y. S. Pu, Aristolochic acid-induced upper tract urothelial carcinoma in taiwan: Clinical characteristics and outcomes. *Int. J. Cancer* **133**, 14–20 (2013).
22. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr., K. W. Kinzler, Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
23. Y. Gui, G. Guo, Y. Huang, X. Hu, A. Tang, S. Gao, R. Wu, C. Chen, X. Li, L. Zhou, M. He, Z. Li, X. Sun, W. Jia, J. Chen, S. Yang, F. Zhou, X. Zhao, S. Wan, R. Ye, C. Liang, Z. Liu, P. Huang, C. Liu, H. Jiang, Y. Wang, H. Zheng, L. Sun, X. Liu, Z. Jiang, D. Feng, J. Chen, S. Wu, J. Zou, Z. Zhang, R. Yang, J. Zhao, C. Xu, W. Yin, Z. Guan, J. Ye, H. Zhang, J. Li, K. Kristiansen, M. L. Nickerson, D. Theodorescu, Y. Li, X. Zhang, S. Li, J. Wang, H. Yang, J. Wang, Z. Cai, Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat. Genet.* **43**, 875–878 (2011).
24. S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, M. R. Stratton, The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* **Chapter 10**, Unit 10.11 (2008).
25. E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, M. R. Stratton, A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
26. E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M. L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordoñez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna, A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, P. J. Campbell, A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
27. W. Lee, Z. Jiang, J. Liu, P. M. Haverty, Y. Guan, J. Stinson, P. Yue, Y. Zhang, K. P. Pant, D. Bhatt, C. Ha, S. Johnson, M. I. Kennemer, S. Mohan, I. Nazarenko, C. Watanabe, A. B. Sparks, D. S. Shames, R. Gentleman, F. J. de Sauvage, H. Stern, A. Pandita, D. G. Ballinger, R. Drmanac, Z. Modrusan, S. Seshagiri, Z. Zhang, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
28. Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
29. Cancer Genome Atlas Research Network, C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, D. A. Levine, Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
30. A. Hartmann, L. Zanardo, T. Bocker-Edmonston, H. Blaszyk, W. Dietmaier, R. Stoehr, J. C. Cheville, K. Junker, W. Wieland, R. Knuechel, J. Rueschoff, F. Hofstaedter, R. Fishel, Frequent microsatellite instability in sporadic tumors of the upper urinary tract. *Cancer Res.* **62**, 6796–6802 (2002).
31. S. S. Hecht, Biochemistry, biology, and carcinogenicity of tobacco-specific *N*-nitrosamines. *Chem. Res. Toxicol.* **11**, 559–603 (1998).
32. B. H. Yun, T. A. Rosenquist, V. Sidorenko, C. R. Iden, C. H. Chen, Y. S. Pu, R. Bonala, F. Johnson, K. G. Dickman, A. P. Grollman, R. J. Turesky, Biomonitoring of aristolactam-DNA adducts in human tissues using ultra-performance liquid chromatography/ion-trap mass spectrometry. *Chem. Res. Toxicol.* **25**, 1119–1131 (2012).
33. J. Fujita, S. K. Srivastava, M. H. Kraus, J. S. Rhim, S. R. Tronick, S. A. Aaronson, Frequency of molecular alterations affecting ras protooncogenes in human urinary tract tumors. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3849–3853 (1985).
34. P. P. Bringuier, M. McCredie, G. Sauter, M. Bilous, J. Stewart, M. J. Mihatsch, P. Kleihues, H. Ohgaki, Carcinomas of the renal pelvis associated with smoking and phenacetin abuse: p53 mutations and polymorphism of carcinogen-metabolising enzymes. *Int. J. Cancer* **79**, 531–536 (1998).
35. J. M. van Oers, E. C. Zwarthoff, I. Rehman, A. R. Azzouzi, O. Cussenot, M. Meuth, F. C. Hamdy, J. W. Catto, *FGFR3* mutations indicate better survival in invasive upper urinary tract and bladder tumours. *Eur. Urol.* **55**, 650–657 (2009).
36. C. Tomasetti, B. Vogelstein, G. Parmigiani, Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1999–2004 (2013).
37. X. Wei, V. Walia, J. C. Lin, J. K. Teer, T. D. Prickett, J. Gartner, S. Davis; NISC Comparative Sequencing Program, K. Stemke-Hale, M. A. Davies, J. E. Gershenwald, W. Robinson, S. Robinson, S. A. Rosenberg, Y. Samuels, Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.* **43**, 442–446 (2011).
38. S. I. Nikolaev, D. Rimoldi, C. Iseli, A. Valsesia, D. Robyr, C. Gehrig, K. Harshman, M. Guipponi, O. Bukach, V. Zoete, O. Michielin, K. Muehlethaler, D. Speiser, J. S. Beckmann, I. Xenarios, T. D. Halazonetis, C. V. Jongeneel, B. J. Stevenson, S. E. Antonarakis, Exome sequencing identifies recurrent somatic *MAP2K1* and *MAP2K2* mutations in melanoma. *Nat. Genet.* **44**, 133–139 (2012).
39. M. S. Stark, S. L. Woods, M. G. Gartside, V. F. Bonazzi, K. Dutton-Regester, L. G. Aoude, D. Chow, C. Sereduk, N. M. Niemi, N. Tang, J. J. Ellis, J. Reid, V. Zismann, S. Tyagi, D. Muzny, I. Newsham,

Y. Wu, J. M. Palmer, T. Pollak, D. Youngkin, B. R. Brooks, C. Lanagan, C. W. Schmidt, B. Kobe, J. P. MacKeigan, H. Yin, K. M. Brown, R. Gibbs, J. Trent, N. K. Hayward, Frequent somatic mutations in *MAP3K5* and *MAP3K9* in metastatic melanoma identified by exome sequencing. *Nat. Genet.* **44**, 165–169 (2012).

40. M. F. Berger, E. Hodis, T. P. Heffernan, Y. L. Deribe, M. S. Lawrence, A. Protopopov, E. Ivanova, I. R. Watson, E. Nickerson, P. Ghosh, H. Zhang, R. Zeid, X. Ren, K. Cibulskis, A. Y. Sivachenko, N. Wagle, A. Sucker, C. Sougnez, R. Onofrio, L. Ambrogio, D. Auclair, T. Fennell, S. L. Carter, Y. Drier, P. Stojanov, M. A. Singer, D. Voet, R. Jing, G. Saksena, J. Barretina, A. H. Ramos, T. J. Pugh, N. Stransky, M. Parkin, W. Winckler, S. Mahan, K. Ardlie, J. Baldwin, J. Wargo, D. Schadendorf, M. Meyerson, S. B. Gabriel, T. R. Golub, S. N. Wagner, E. S. Lander, G. Getz, L. Chin, L. A. Garraway, Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–506 (2012).
41. E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J. P. Theurillat, E. Nickerson, D. Auclair, L. Li, C. Place, D. Dicara, A. H. Ramos, M. S. Lawrence, K. Cibulskis, A. Sivachenko, D. Voet, G. Saksena, N. Stransky, R. C. Onofrio, W. Winckler, K. Ardlie, N. Wagle, J. Wargo, K. Chong, D. L. Morton, K. Stemke-Hale, G. Chen, M. Noble, M. Meyerson, J. E. Ladbury, M. A. Davies, J. E. Gershenwald, S. N. Wagner, D. S. Hoon, D. Schadendorf, E. S. Lander, S. B. Gabriel, G. Getz, L. A. Garraway, L. Chin, A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
42. M. Krauthammer, Y. Kong, B. H. Ha, P. Evans, A. Bacchiocchi, J. P. McCusker, E. Cheng, M. J. Davis, G. Goh, M. Choi, S. Ariyan, D. Narayan, K. Dutton-Regester, A. Capatana, E. C. Holman, M. Bosenberg, M. Sznol, H. M. Kluger, D. E. Brash, D. F. Stern, M. A. Materin, R. S. Lo, S. Mane, S. Ma, K. K. Kidd, N. K. Hayward, R. P. Lifton, J. Schlessinger, T. J. Boggon, R. Halaban, Exome sequencing identifies recurrent somatic *RAC1* mutations in melanoma. *Nat. Genet.* **44**, 1006–1014 (2012).
43. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Tobacco Smoke and Involuntary Smoking* (World Health Organization–International Agency for Research on Cancer, Lyon, 2004).
44. P. Liu, C. Morrison, L. Wang, D. Xiong, P. Vedell, P. Cui, X. Hua, F. Ding, Y. Lu, M. James, J. D. Ebben, H. Xu, A. A. Adjei, K. Head, J. W. Andrae, M. R. Tschannen, H. Jacob, J. Pan, Q. Zhang, F. Van den Bergh, H. Xiao, K. C. Lo, J. Patel, T. Richmond, M. A. Watt, T. Albert, R. Selzer, M. Anderson, J. Wang, Y. Wang, S. Starnes, P. Yang, M. You, Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* **33**, 1270–1276 (2012).
45. C. M. Rudin, S. Durinck, E. W. Stawiski, J. T. Poirier, Z. Modrusan, D. S. Shames, E. A. Bergbower, Y. Guan, J. Shin, J. Guillory, C. S. Rivers, C. K. Foo, D. Bhatt, J. Stinson, F. Gnad, P. M. Haverty, R. Gentleman, S. Chaudhuri, V. Janakiraman, B. S. Jaiswal, C. Parikh, W. Yuan, Z. Zhang, H. Koeppen, T. D. Wu, H. M. Stern, R. L. Yauch, K. E. Huffman, D. D. Paskulin, P. B. Illei, M. Varella-Garcia, A. F. Gazdar, F. J. de Sauvage, R. Bourgon, J. D. Minna, M. V. Brock, S. Seshagiri, Comprehensive genomic analysis identifies *SOX2* as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.* **44**, 1111–1116 (2012).
46. R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, K. Chen, J. Walker, S. McDonald, R. Bose, D. Ornitz, D. Xiong, M. You, D. J. Dooling, M. Watson, E. R. Mardis, R. K. Wilson, Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).
47. M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansén, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Jänne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, M. Meyerson, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
48. D. Xiong, G. Li, K. Li, Q. Xu, Z. Pan, F. Ding, P. Vedell, P. Liu, P. Cui, X. Hua, H. Jiang, Y. Yin, Z. Zhu, X. Li, B. Zhang, D. Ma, Y. Wang, M. You, Exome sequencing identifies MXRA5 as a novel cancer gene frequently mutated in non–small cell lung carcinoma from Chinese patients. *Carcinogenesis* **33**, 1797–1805 (2012).
49. N. Agrawal, M. J. Frederick, C. R. Pickering, C. Bettegowda, K. Chang, R. J. Li, C. Fakhry, T. X. Xie, J. Zhang, J. Wang, N. Zhang, A. K. El-Naggar, S. A. Jasser, J. N. Weinstein, L. Treviño, J. A. Drummond, D. M. Muzny, Y. Wu, L. D. Wood, R. H. Hruban, W. H. Westra, W. M. Koch, J. A. Califano, R. A. Gibbs, D. Sidransky, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, D. A. Wheeler, K. W. Kinzler, J. N. Myers, Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333**, 1154–1157 (2011).
50. S. L. Poon, S.-T. Pang, J. R. McPherson, W. Yu, K. K. Huang, P. Guan, W.-H. Weng, E. Y. Siew, Y. Liu, H. L. Heng, S. C. Chong, A. Gan, S. T. Tay, W. K. Lim, I. Cutcutache, D. Huang, L. D. Ler, M.-L. Nairismägi, M. H. Lee, Y.-H. Chang, K.-J. Yu, W. Chan-on, B.-K. Li, Y.-F. Yuan, C.-N. Qian, K.-F. Ng, C.-F. Wu, C.-L. Hsu, R. M. Bunte, M. R. Stratton, P. A. Futreal, W.-K. Sung, C.-K. Chuang, C. K. Ong, S. G. Rozen, P. Tan, B. T. Teh, Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
51. A. K. Goodenough, H. A. Schut, R. J. Turesky, Novel LC-ESI/MS/MS$^n$ method for the characterization and quantification of 2′-deoxyguanosine adducts of the dietary carcinogen 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine by 2-D linear quadrupole ion trap mass spectrometry. *Chem. Res. Toxicol.* **20**, 263–276 (2007).
52. C. Bettegowda, N. Agrawal, Y. Jiao, M. Sausen, L. D. Wood, R. H. Hruban, F. J. Rodriguez, D. P. Cahill, R. McLendon, G. Riggins, V. E. Velculescu, S. M. Oba-Shinjo, S. K. Marie, B. Vogelstein, D. Bigner, H. Yan, N. Papadopoulos, K. W. Kinzler, Mutations in *CIC* and *FUBP1* contribute to human oligodendroglioma. *Science* **333**, 1453–1455 (2011).
53. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
54. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
55. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA; http://evs.gs.washington.edu/EVS/ [accessed April 2012].
56. T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
57. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

**Acknowledgments:** We thank L. Dobbyn, J. Ptak, J. Schaefer, N. Silliman, D. Singh, and G. Mihalyne for excellent technical and quality control assistance. **Funding:** Supported by The Virginia and D. K. Ludwig Fund for Cancer Research; Commonwealth Foundation; NIH grants CA57345 (K.W.K.), ES004068 (A.P.G.), and ES019564 (R.J.T.); Taiwan National Science Council (101-2314-B-002-027-MY3); Department of Health, Taiwan (DOH101-TD-C-111-001); and Henry and Marsha Laufer. T.A.R. was the recipient of a Zickler Translational Research Scholar Award funded by the Zickler Family Foundation. **Author contributions:** M.L.H., B.V., N.P., A.P.G., K.W.K., Y.-S.P., and T.A.R. designed the study; C.-H.C. and K.G.D. collected and analyzed the UTUC samples; B.H.Y., R.J.T., and V.S.S. determined AL-DNA adduct levels; M.L.H., J.H., and N.P. performed genomic sequencing; M.L.H., N.N., C.D., R.K., M.M., N.P., K.W.K., and T.A.R. analyzed the genetic data; M.L.H., A.P.G., K.W.K., and T.A.R. wrote draft manuscripts. All authors contributed to the final version of the paper. **Competing interests:** Under agreements between the Johns Hopkins University, Genzyme, Exact Sciences, Inostics, Qiagen, Invitrogen, and Personal Genome Diagnostics, N.P., B.V., and K.W.K. are entitled to a share of the royalties received by the University on sales of products related to genes reported in this manuscript. N.P., B.V., and K.W.K. are co-founders of Inostics and Personal Genome Diagnostics, are members of their Scientific Advisory Boards, and own Inostics and Personal Genome Diagnostics stock, which is subject to certain restrictions under Johns Hopkins University policy. **Data and materials availability:** The list of somatic mutations identified in this study is available in table S3. The list of COSMIC mutations used in this study is available in table S6.

Submitted 22 March 2013
Accepted 11 July 2013
Published 7 August 2013
10.1126/scitranslmed.3006200

**Citation:** M. L. Hoang, C.-H. Chen, V. S. Sidorenko, J. He, K. G. Dickman, B. H. Yun, M. Moriya, N. Niknafs, C. Douville, R. Karchin, R. J. Turesky, Y.-S. Pu, B. Vogelstein, N. Papadopoulos, A. P. Grollman, K. W. Kinzler, T. A. Rosenquist, Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5**, 197ra102 (2013).