nature genetics

Somatic *ERCC2* mutations are associated with a distinct genomic signature in urothelial tumors

Jaegil Kim[1,10], Kent W Mouw[2,3,10], Paz Polak[1,3–5,10], Lior Z Braunstein[1,3], Atanas Kamburov[1,3–5], Grace Tiao[1], David J Kwiatkowski[3,6], Jonathan E Rosenberg[7], Eliezer M Van Allen[1,3,8], Alan D D'Andrea[2,3,9] & Gad Getz[1,3–5]

Alterations in DNA repair pathways are common in tumors and can result in characteristic mutational signatures; however, a specific mutational signature associated with somatic alterations in the nucleotide- excision repair (NER) pathway has not yet been identified. Here we examine the mutational processes operating in urothelial cancer, a tumor type in which the core NER gene *ERCC2* is significantly mutated. Analysis of three independent urothelial tumor cohorts demonstrates a strong association between somatic *ERCC2* mutations and the activity of a mutational signature characterized by a broad spectrum of base changes. In addition, we note an association between the activity of this signature and smoking that is independent of *ERCC2* mutation status, providing genomic evidence of tobacco-related mutagenesis in urothelial cancer. Together, these analyses identify an NER-related mutational signature and highlight the related roles of DNA damage and subsequent DNA repair in shaping tumor mutational landscape.

Cells are continually exposed to both exogenous and endogenous sources of DNA damage, and multiple DNA repair pathways have evolved to repair a variety of DNA lesions. However, many tumors are functionally deficient in one or more DNA repair pathways[1–3]. The somatic mutational landscape of tumor cells reflects the cumulative activity of discrete mutational processes operating across the lifetime of each cell, and loss of DNA repair fidelity can augment the effect of these processes and lead to increased somatic mutation rates.

Mutational signatures are patterns of base changes associated with specific mutational processes operating in tumor cells. Recently, non-negative matrix factorization (NMF) methods have been applied to discover and characterize mutational signatures across multiple tumor types[4,5]. Dozens of mutational signatures have been identified, including several that have been linked to specific DNA-damaging agents or DNA repair defects[6]. Mutational signatures associated with deficiencies in the homologous recombination and mismatch-repair pathways have recently been characterized, but mutational signatures associated with deficiencies in other DNA repair pathways have not yet been identified.

The NER pathway is a highly conserved DNA repair pathway that removes bulky intrastrand adducts created by agents such as UV radiation and certain chemicals, including several commonly used chemotherapy agents[7]. Somatic mutations in NER pathway genes occur sporadically across cancer types, but recurrent mutations in specific NER pathway genes are uncommon[8]. One notable exception is the *ERCC2* gene, which encodes a DNA helicase that has a central role in the NER pathway, unwinding the DNA duplex adjacent to a site of damage[9,10]. Recurrent somatic *ERCC2* mutations have been identified in 6–18% of urothelial tumors in studies published by The Cancer Genome Atlas (TCGA) and others[11–14].

Tumors of the urothelial tract and bladder account for nearly 75,000 new cancer cases each year in the United States and are associated with exposure to tobacco, chemicals, and certain infectious agents[15,16]. Many of these carcinogens are known to damage DNA through the formation of bulky intrastrand adducts[17–19], and several studies have demonstrated an increased risk of bladder cancer in individuals with polymorphisms in *ERCC2* or other NER pathway genes[20,21]. Similar to other carcinogen-associated tumors, most urothelial tumors have a high somatic mutation burden. *ERCC2*-mutated urothelial tumors have a higher overall mutation burden than tumors with wild-type *ERCC2* but have a lower fraction of C>G mutations[11]. Despite the known association between smoking and urothelial cancer, a tobacco-associated mutational signature has not been identified in urothelial tumors.

To more fully characterize the mutational processes operating in urothelial cancer, we performed mutational signature analysis in three independent urothelial tumor cohorts. Our analysis identified four operating mutational signatures, including one signature for which an etiology had not previously been described. Unbiased enrichment analysis identified *ERCC2* as the gene that, when mutated, was most strongly associated with the activity of this signature in all three cohorts. Furthermore, we find that activity of the signature is associated with smoking history, making this the first description of

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [6]Division of Pulmonary Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [7]Genitourinary Oncology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [9]Center for DNA Damage and Repair, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to G.G. (gadgetz@broadinstitute.org).

Received 8 July 2015; accepted 1 April 2016; published online 25 April 2016; doi:10.1038/ng.3557

a tobacco-related mutational signature in urothelial cancer. Together, these findings identify the first NER-related mutational signature, to our knowledge, and underscore the importance of both exposure to DNA-damaging agents and operation of DNA repair pathways in the activities of mutational signatures.

RESULTS

Mutational signature analysis of the TCGA-130 cohort

To understand the DNA damage and repair processes operating in urothelial tumors, we performed mutational signature analysis of 130 muscle-invasive urothelial tumors from the TCGA-130 cohort (**Fig. 1a** and **Supplementary Table 1**). We applied a Bayesian variant of the NMF algorithm to mutation counts, stratified by 96 trinucleotide mutational contexts, to infer (i) the number of operating mutational processes, (ii) their signatures (96 normalized weights per process), and (iii) the activity of each signature in every tumor (the estimated number of mutations associated with each signature) (Online Methods)[5,22].

Our analysis identified four independent mutational signatures in the TCGA-130 cohort (**Fig. 1b** and **Supplementary Tables 2** and **3**), and although our analysis methods are not identical to those applied by the Sanger Institute our signatures matched four of the previously identified Sanger signatures (cosine similarities between 0.86 and 0.95), which are described in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (**Supplementary Fig. 1** and **Supplementary Table 4**)[4]. Two of the signatures, characterized by C>T transitions and C>G transversions at TC[A/T] motifs (where the mutated C is preceded by T and followed by A or T), occur in multiple tumor types and are attributed to APOBEC-mediated mutagenesis (denoted as APOBEC1 and APOBEC2 in **Fig. 1b** and corresponding to COSMIC signatures 13 and 2, respectively)[4,23,24]. A third signature, characterized by C>T transitions at CpG dinucleotides, is found in all tumor types and is thought to result from age-related accumulation of 5-methylcytosine deamination events (C>T CpG in **Fig. 1b**; COSMIC signature 1). Finally, a fourth signature was identified that closely resembles COSMIC signature 5 (cosine similarity of 0.90; denoted as signature 5* in **Fig. 1b** and **Supplementary Fig. 1**). COSMIC signature 5 is characterized by a broad spectrum of base changes and is present in all tumor types; an etiology has not yet been described.

Signature 5* activity is associated with *ERCC2* mutations

To further characterize signature 5*, we performed signature enrichment analysis to identify genes that, when mutated, were associated with increased activity of signature 5* (Online Methods). For each of the 283 genes that had a non-silent mutation in >5% of samples across the TCGA-130 cohort, we compared the activity of signature 5* in tumors that carried a non-silent mutation in the gene and tumors that did not. To ensure that increased signature 5* activity did not reflect an increase in overall mutation burden, we assessed the significance level using a permutation-based method that controls the overall mutation burden in each sample (Online Methods). *ERCC2* was the only significant gene (Benjamini–Hochberg false discovery rate (FDR) $q = 8.6 \times 10^{-3}$, $P = 3 \times 10^{-5}$; **Fig. 2** and **Supplementary Fig. 2**). Overall, 16 of the 130 tumors had a non-silent *ERCC2* mutation, and these tumors had a median of 135 signature 5* mutations as compared to 40 in tumors without a non-silent *ERCC2* mutation (an increase of 95 mutations in the median activity of signature 5*) (**Fig. 3a**).

Validation in two independent cohorts of urothelial tumors

To validate these findings, we performed similar analyses on two independent cohorts. The first cohort included 50 muscle-invasive urothelial tumors recently analyzed by Van Allen *et al.* (DFCI/MSK-50 cohort) (**Supplementary Table 1**)[12]. This cohort comprises patients treated with neoadjuvant cisplatin-based chemotherapy and contains an equal number of cisplatin responders and non-responders. Bayesian NMF analysis yielded four mutational signatures that closely resembled the signatures identified in the TCGA-130 cohort (cosine similarities of 0.93–0.99; **Supplementary Figs. 1** and **3**, and **Supplementary Table 5**). Repeating the gene mutation enrichment analysis for signature 5* activity identified three significant genes ($q \leq 0.1$), with *ERCC2* being the most significant ($q = 0.042$, $P = 1.9 \times 10^{-4}$; **Fig. 2** and **Supplementary Figs. 2** and **4**). Nine of the 50 tumors had a non-silent mutation in *ERCC2*, and these tumors had a median of 220 signature 5* mutations as compared to 32 in tumors lacking non-silent *ERCC2* mutations (an increase of 188).

The second validation cohort comprised 99 urothelial tumors (62 muscle invasive and 37 non muscle invasive) recently reported by Guo *et al.* (BGI-99) (**Supplementary Table 1**)[13]. As in the previous two cohorts, our analysis identified four mutational signatures (**Supplementary Fig. 5**). The first two resembled the two APOBEC-associated signatures (cosine similarities of 0.96 and 0.80 with COSMIC signatures 2 and 13, respectively), but the third signature was not observed in the previous cohorts and was dominated by T>A mutations. This signature is most similar to COSMIC signature 22 (cosine similarity of 0.96) and has been linked to exposure to aristolochic acid, an ingredient in some food supplements that are most commonly used in Asian countries[25]. Indeed, consumption of aristolochic acid has been associated with increased risk of urothelial cancers[26–28]. The fourth signature in this cohort seems to be a

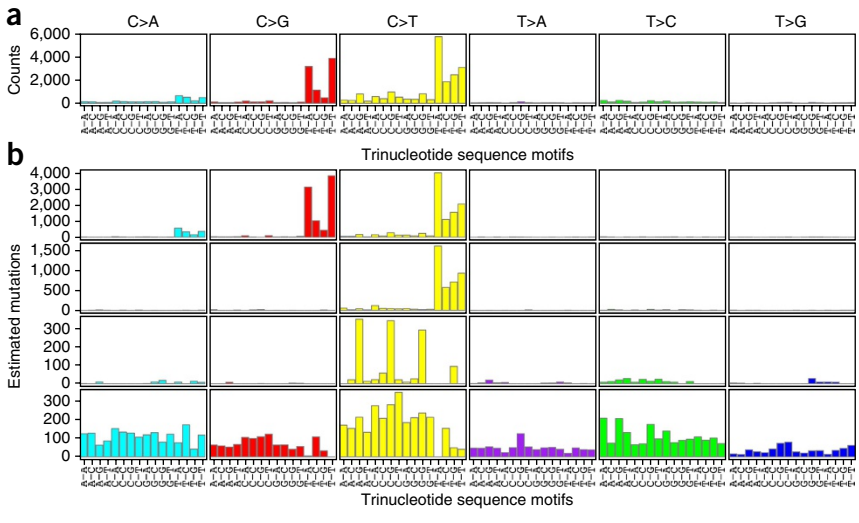

Figure 1 Mutational signature analysis of 130 TCGA muscle-invasive urothelial tumors (TCGA-130 cohort). (**a**) The spectrum of base changes identified in the TCGA-130 cohort displayed for mutated pyrimidines and adjacent 3′ and 5′ bases. (**b**) A Bayesian NMF algorithm was applied to identify signatures from the matrix of mutation counts across tumors. Four distinct mutational signatures were identified.

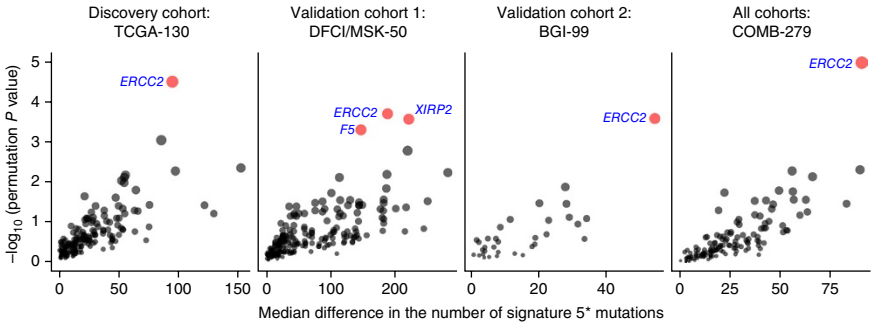

Figure 2 Mutation enrichment analysis identifies an association between somatic *ERCC2* mutations and activity of signature 5* in a discovery cohort, two validation cohorts, and the combined cohort. For genes mutated in >5% of samples in each cohort, the number of mutations attributed to signature 5* was compared in tumors with wild-type versus mutated copies of the gene while controlling for overall mutation burden. Genes with FDR $q < 0.1$ are highlighted in red. *ERCC2* was the only gene that was significant in each of the cohorts. COMB-279 refers to the combined cohort (TCGA-130 + DFCI/MSK-50 + BGI-99).

superposition of the two other signatures identified in the previous cohorts, C>T CpG and signature 5*, with the lack of separation possibly due to insufficient resolution given the lower overall mutation burden in this cohort. As in the other cohorts, tumors with a non-silent mutation in *ERCC2* had increased activity of the fourth signature, which includes signature 5* (ten tumors with a non-silent *ERCC2* mutation and a median increase of 55 mutations per sample; $q = 0.012$, $P = 2.5 \times 10^{-4}$; **Fig. 2** and **Supplementary Figs. 2** and **4**).

Finally, we repeated the analysis for all 279 tumors across the three cohorts (COMB-279 cohort). Among the 35 tumors with a non-silent *ERCC2* mutation, the median signature 5* activity was increased by 91 mutations in comparison to tumors with wild-type *ERCC2* (124 versus 33), providing the strongest statistical evidence for the association between *ERCC2* mutation and signature 5* activity ($q = 1.6 \times 10^{-3}$, $P = 1.0 \times 10^{-5}$; **Fig. 2** and **Supplementary Figs. 2** and **4**). Together, these data strongly suggest that, although signature 5* activity is present in tumors both wild type and mutant for *ERCC2*, somatic *ERCC2* mutations are associated with a significant increase in signature 5* activity.

To further characterize the association between *ERCC2* mutational status and signature 5* activity, we performed unsupervised clustering of tumors based on signature 5* activity (in 96 trinucleotide mutational contexts). The combined cohort (COMB-279) segregated into two clusters of 222 and 57 tumors, with 25 of the 35 *ERCC2*-mutated tumors in the second cluster ($P = 1.7 \times 10^{-12}$, Fisher's exact test; **Supplementary Fig. 6a**). Repeating the analysis using the 242 muscle-invasive tumors across cohorts (COMB-MI-242) yielded a more significant association between clusters and *ERCC2* mutations ($P = 4.4 \times 10^{-14}$; **Supplementary Fig. 6b**). Although *ERCC2* mutations are associated with higher overall mutation burden[11–13], segregation was not driven by this higher burden, as *ERCC2*-mutated tumors segregated less strongly when clustering was performed using the total number of single-nucleotide variants (SNVs) ($P_{\text{COMB-279}} = 0.1$ and $P_{\text{COMB-MI-242}} = 0.008$; **Supplementary Fig. 6c,d**).

All but one of the 35 non-silent *ERCC2* mutations across the cohorts were missense mutations, and most of these (25 of 34) mapped within or adjacent to (±10 amino acids) the conserved helicase motifs, suggesting that the mutations may have an impact on ERCC2 protein function (**Supplementary Fig. 7a**). Supporting this hypothesis, the mutations mapping to helicase motifs were associated with higher signature 5* activity than mutations mapping elsewhere in the protein (median number of signature 5* mutations of 134 versus 96, $P = 0.037$). To assess the spatial relationship of protein residues affected by mutation, we used CLUMPS, a novel algorithm for assessing spatial clustering of altered residues within three-dimensional protein structures, and found that the residues corresponding to *ERCC2* mutations were significantly clustered ($P = 0.0026$), further suggesting that the mutations have a functional role (Online Methods and **Supplementary Fig. 7b**)[29].

For each of the three cohorts analyzed here, *ERCC2*-mutated tumors have been shown to have greater overall mutation burden than tumors with wild-type *ERCC2* (refs. 11–13). We asked whether this higher burden was due solely to increased signature 5* activity or whether the activities of other signatures were also increased. Indeed, activity of the APOBEC2 signature was also higher in *ERCC2*-mutated tumors than in tumors with wild-type *ERCC2* in the TCGA-130 cohort (39 versus 12 APOBEC2 mutations, $P = 0.004$; **Fig. 3b**); however, unlike the association between *ERCC2* mutation and signature 5*, the association of *ERCC2* mutation with the APOBEC2 signature was not

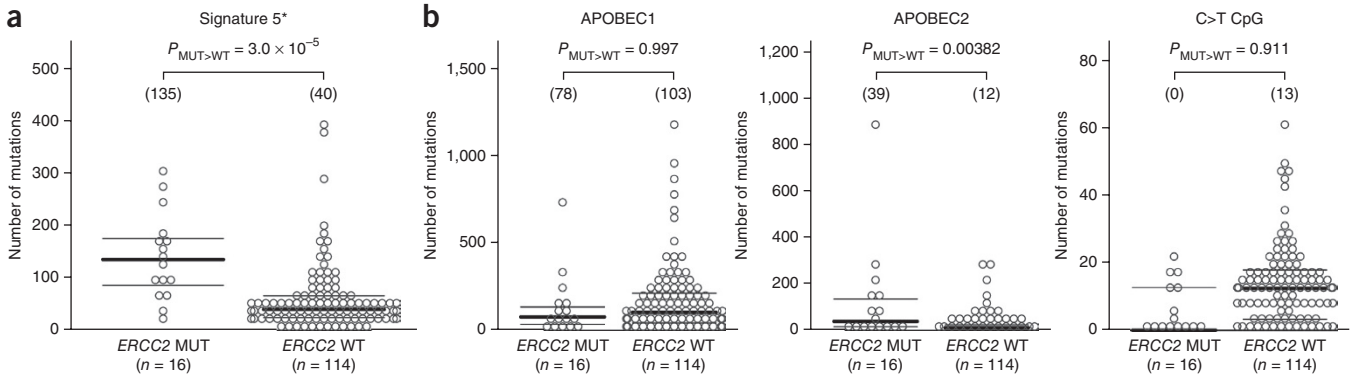

Figure 3 Comparison of signature activities in tumors with wild-type versus mutant *ERCC2* in the TCGA-130 cohort. (**a**) The estimated number of signature 5* mutations was significantly higher in tumors with mutated (MUT) *ERCC2* than in tumors with wild-type (WT) *ERCC2*. (**b**) Estimated numbers of mutations attributed to the three other mutational signatures identified in the TCGA-130 cohort. The median estimated number of mutations is shown in parentheses, and *P* values were computed using a one-tailed permutation test. Horizontal lines represent median (bold) and upper and lower quartile values.

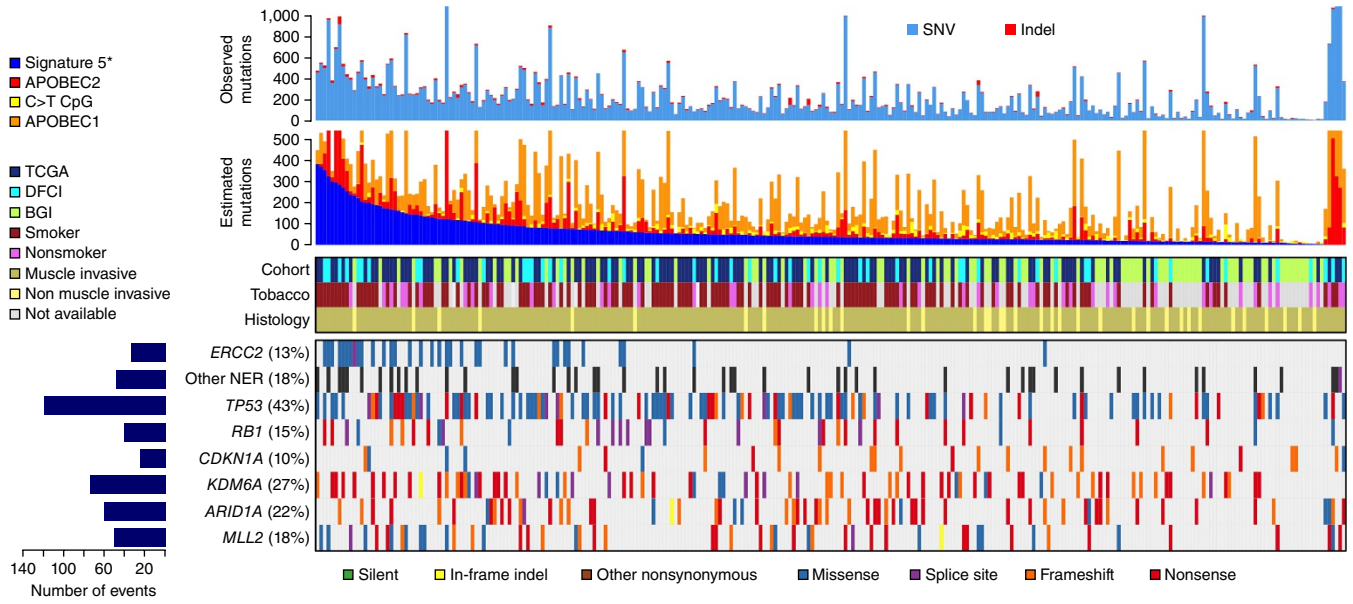

Figure 4 Overall mutation rates, mutational signature contributions, and mutational status of *ERCC2* and other genes of interest in the combined cohort (TCGA-130 + DFCI/MSK-50 + BGI-99). Each column represents a tumor. Overall mutation burden is shown at the top, followed by the estimated contribution of each of the four mutational signatures to the overall mutation burden (samples are arranged in order of decreasing signature 5* activity), cohort, smoking status, and histology (muscle invasive versus non muscle invasive). In the bottom half of the figure, the mutational status of *ERCC2* and other genes of interest is shown, with events color-coded by type of mutation. Somatic events in non-*ERCC2* NER pathway genes are collapsed into a single track (see **Supplementary Fig. 10** for an expanded list of NER pathway genes) and are followed by other significantly mutated genes in urothelial cancer (*TP53*, *RB1*, etc.).

significant after correcting for multiple testing ($q = 0.54$). A similar association between *ERCC2* mutation and the APOBEC2 signature was seen in the other two cohorts, but this association was only statistically significant in the combined (COMB-279) cohort ($q = 0.0016$; **Supplementary Figs. 8** and **9**). There was no increase in activity of the APOBEC1 (78 versus 103 mutations in TCGA-130, $P = 0.99$) or C>T CpG (0 versus 13, $P = 0.91$) signature in tumors with mutant versus wild-type *ERCC2* in any of the cohorts (**Fig. 3b** and **Supplementary Fig. 8**). These results demonstrate that the higher overall mutation burden in *ERCC2*-mutated tumors is due primarily to increased activity of signature 5*, with an additional smaller contribution from the APOBEC2 signature.

Non-*ERCC2* NER mutations and signature 5* activity

Despite the strong association between signature 5* activity and *ERCC2* mutational status, several tumors with high signature 5* activity lacked a somatic *ERCC2* mutation. In these cases, we hypothesized that other somatic or germline NER pathway alterations might contribute to signature 5* activity. Somatic mutations in other NER pathway genes are less common in urothelial tumors, and there was no statistically significant association between signature 5* activity and mutations in any individual NER gene or the pathway as a whole (when *ERCC2* was excluded) (**Fig. 4** and **Supplementary Fig. 10**). However, anecdotally, of the 20 tumors with wild-type *ERCC2* showing the highest signature 5* activity, 6 had a mutation in a different gene in the NER pathway. In addition, germline data were available for the TCGA-130 and DFCI/MSK-50 cohorts, and there was a trend toward an association between rare (<2% frequency in the cohorts) NER germline variants and signature 5* activity in cases with wild-type *ERCC2* (19 of the 32 tumors with wild-type *ERCC2* showing the highest signature 5* activity had an NER germline variant in comparison to only 54 of the remaining 123 tumors with wild-type *ERCC2*, $P = 0.086$; Online Methods and **Supplementary Fig. 11a**). Moreover, four specific NER germline alleles were enriched ($q < 0.1$) in tumors with wild-type *ERCC2* showing high signature 5* activity, and three of the four are predicted to be functionally deleterious (**Supplementary Fig. 11b**)[30]. However, additional studies in larger cohorts will be needed to further characterize the potential contribution of non-*ERCC2* somatic and germline NER alterations to signature 5* activity.

Smoking is associated with signature 5* activity

Given the known association between smoking and urothelial cancer, we attempted to identify evidence of tobacco exposure in the mutational signatures of urothelial tumors. Data on smoking status were available for the TCGA-130 and DFCI/MSK-50 cohorts, and these were therefore analyzed together. There was no difference in overall mutation burden in tumors from patients with any smoking history ('smokers') versus those with no smoking history ('nonsmokers') ($P = 0.27$, Wilcoxon rank-sum test; **Fig. 5a**). However, the activity of signature 5* was significantly higher in smokers than in nonsmokers (median number of signature 5* mutations, 49 versus 33, $P = 0.009$; **Fig. 5b**), although the effect size for smoking was modest when compared to that of an *ERCC2* mutation (**Fig. 5c**). There were no differences in signature 5* activity between current and former smokers; however, there was a correlation between smoking intensity (pack-years) and signature 5* activity in *ERCC2*-mutated cases ($P = 0.01$; **Supplementary Fig. 12**). There were no differences in other mutational signatures in smokers versus nonsmokers (**Supplementary Fig. 13**).

Although an association between smoking and COSMIC signature 5 has previously been noted in lung adenocarcinoma, a different and more common smoking-related signature characterized by frequent C>A transversions (COSMIC signature 4) was not identified

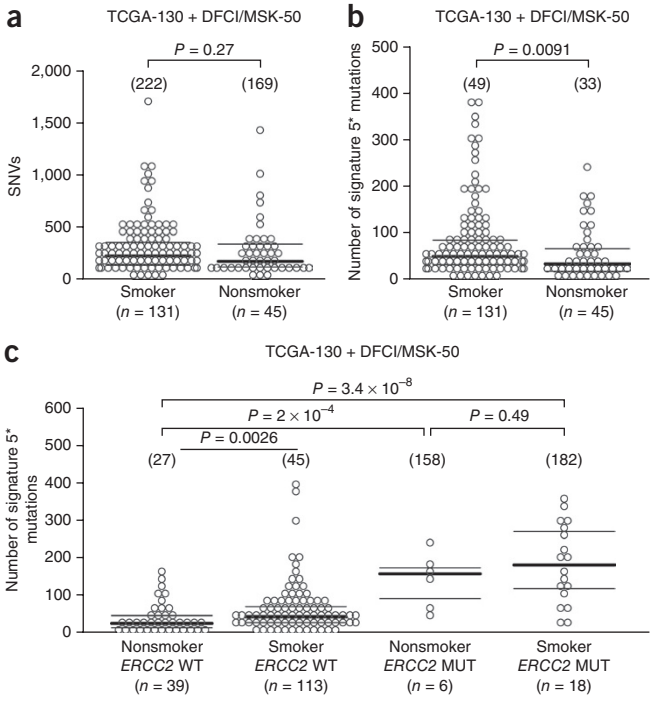

Figure 5 Effect of smoking and *ERCC2* mutational status on signature 5* activity. (**a**) There was no significant difference in the total number of mutations (SNVs) in smokers versus nonsmokers in the combined TCGA-130 + DFCI/MSK-50 cohort. (**b**) The estimated number of signature 5* mutations was significantly higher in smokers than in nonsmokers. (**c**) Among patients harboring tumors with wild-type *ERCC2*, the number of signature 5* mutations was significantly higher in smokers than in nonsmokers, whereas smoking was not associated with a further increase in signature 5* activity among patients with *ERCC2*-mutated tumors. The association between smoking and signature 5* activity is not as strong as the association between *ERCC2* mutation and signature 5* activity. The median number of mutations is shown in parentheses, and *P* values were calculated using the Wilcoxon rank-sum test.

in any of the urothelial cohorts analyzed here[4,31]. Given the association of signature 5* activity with smoking, we explored whether COSMIC signature 4 contributes to signature 5*, with these processes not separated by NMF owing to insufficient power. To test this hypothesis, we attempted to separate signature 5* mutations into contributions from COSMIC signatures 4 and 5 (Online Methods). This analysis confirmed the high similarity of signature 5* and COSMIC signature 5 and demonstrated that the smoking-related difference in signature 5* activity is indeed driven by a difference in activity of COSMIC signature 5 and not COSMIC signature 4 (**Supplementary Fig. 14**).

Several mutational signatures exhibit an asymmetric pattern of mutations on the transcribed versus non-transcribed DNA strand, a phenomenon that is attributed to the increased rate of high-fidelity repair of the transcribed strand by the transcription-coupled repair subpathway of NER[7,32–35]. To determine whether signature 5* exhibits transcriptional strand bias, we repeated the Bayesian NMF analysis using 192 mutational contexts (instead of 96) to consider mutations on the transcribed and non-transcribed strands independently (Online Methods and **Supplementary Fig. 15**). Signature 5* exhibited strand asymmetry in several contexts, including a bias for T>C transitions on the transcribed strand, as described for COSMIC signature 5 (ref. 4). In addition, a bias for C>A transversions on the transcribed strand (similar to COSMIC signature 4) was also observed and may arise from the decreased rate of repair of tobacco-induced guanine damage on the non-transcribed strand[6].

The activity of a mutational signature depends both on the potency of the mutagenic process and the length of time over which it operates. Recently, activity of COSMIC signatures 1 and 5 was found to be correlated with patient age, suggesting that the underlying mutational processes are active across the lifetime of somatic cells[36]. However, no association between age and COSMIC signature 5 activity was found in urothelial cancer, indicating that other factors drive signature 5 activity. Independent analysis of the TCGA-130 and DFCI/MSK-50 cohorts (the two cohorts with available age data in our study) also failed to identify an association between age and signature 5* activity ($P = 0.65$; **Supplementary Fig. 16**). Similarly, on multivariate regression analysis, *ERCC2* mutational status ($P = 3.5 \times 10^{-14}$) and smoking ($P = 0.038$) were significantly associated with signature 5* activity, whereas age ($P = 0.60$) and sex ($P = 0.48$) were not.

Somatic *ERCC2* mutations drive signature 5* activity

To further investigate the factors influencing signature 5* activity, we used ABSOLUTE to estimate the cancer cell fraction (CCF) of each mutation in the 126 tumors from the TCGA-130 cohort for which allelic copy number data were available (Online Methods). Sixteen of the 126 tumors (13%) had a somatic *ERCC2* mutation, and all 16 mutations were heterozygous. Eleven of the 16 mutations were clonal (defined as Pr(CCF ≥ 0.95) >0.5) and 5 were subclonal. We reasoned that, if *ERCC2* mutations are responsible for increasing the number of signature 5* mutations (rather than just being associated with higher signature 5* activity), then tumors with clonal *ERCC2* mutations would have a higher ratio of clonal to subclonal signature 5* mutations than tumors with subclonal *ERCC2* mutations. Supporting this hypothesis, we found that clonal signature 5* mutations were enriched in tumors with clonal mutations of *ERCC2* (clonal/subclonal ratio ~5, $P = 0.0098$, pairwise Mann–Whitney test) but not in tumors with subclonal *ERCC2* mutations (clonal/subclonal ratio ~1.1, $P = 0.81$) or with wild-type *ERCC2* (clonal/subclonal ratio ~1.9, $P = 0.49$; **Fig. 6** and **Supplementary Fig. 17**). Overall, these data suggest that somatic

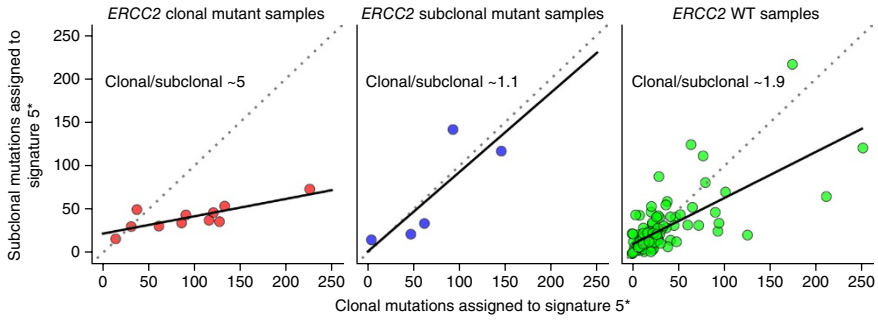

Figure 6 Association between clonality of *ERCC2* mutations and clonality of signature 5* mutations. For tumors with a clonal *ERCC2* mutation (defined by Pr(CCF ≥ 0.95) >0.5) (left), the majority of signature 5* mutations are clonal (clonal/subclonal ratio ~5). For tumors with a subclonal *ERCC2* mutation (middle) or wild-type *ERCC2* (right), the ratio of clonal to subclonal signature 5* mutations was much lower (clonal/subclonal ratio ~1.1 and ~1.9, respectively).

ERCC2 mutations are often early events in tumorigenesis and drive signature 5* activity.

Signature 5* activity and cisplatin response

Platinum-based therapies are widely used in the treatment of urothelial cancers, but individual patients vary in their response to treatment. Therefore, predictive biomarkers are needed to guide therapy. We recently showed that *ERCC2* mutations are enriched in urothelial tumors responsive to cisplatin-based chemotherapy, and other studies have identified additional genetic alterations that characterize cisplatin-responsive tumors[37–40]. Of the cohorts analyzed here, only the DFCI/MSK-50 cohort had cisplatin response data available, and there was significantly higher signature 5* activity in the 25 cisplatin responders than in the 25 non-responders ($P = 0.027$; **Supplementary Fig. 18**); however, signature 5* activity was not associated with cisplatin response in cases with wild-type *ERCC2* ($P = 0.51$). Additional studies in larger cohorts will be needed to determine whether signature 5* activity can be used to predict platinum response in urothelial cancer.

DISCUSSION

Here we identify and validate an association between somatic non-silent mutations in *ERCC2* and activity of a specific mutational signature in three independent urothelial tumor cohorts. The signature is very similar to COSMIC signature 5 (although detected using a slightly different methodology applied to different data sets and hence called signature 5* here; **Supplementary Fig. 1**) and is characterized by a broad pattern of base substitutions[4]. Other signatures identified in our analysis also resemble described signatures, and all have previously been linked to specific underlying mutational processes[11,24,36].

Urothelial cancer is unique in that it is the only known tumor type in which the core NER gene *ERCC2* is significantly mutated[8]. However, COSMIC signature 5 activity has been identified in all tumor types characterized thus far. Therefore, it is unlikely that *ERCC2* mutations are solely responsible for signature 5* activity across tumor types. Instead, signature 5* (and COSMIC signature 5) may reflect the footprint of lower-fidelity DNA repair pathways such as translesion synthesis that normally operate in parallel with high-fidelity repair pathways such as NER and are upregulated when high-fidelity repair is compromised[41,42]. In urothelial cancer, somatic *ERCC2* mutations seem to be the most common genetic event driving upregulation of lower-fidelity repair pathways and signature 5* activity, whereas, in other tumor types, signature 5* activity may result from other genetic or environmental factors that result in increased activity of lower-fidelity repair pathways. Given that recurrent *ERCC2* mutations seem to be unique to urothelial cancer and are often early events in tumorigenesis, additional efforts to understand the role of *ERCC2* in bladder tumor biology may provide important insights.

In addition to the association with *ERCC2* mutational status, we also found that signature 5* activity was increased in smokers, although the effect from smoking was modest relative to the effect from an *ERCC2* mutation. Tobacco exposure is a known risk factor for urothelial cancer; however, unlike other tobacco-related tumors (such as lung squamous cell, lung adenocarcinoma, and head and neck squamous cell cancers), an association between smoking and the activity of a specific mutational signature had not previously been described in urothelial tumors. Here we noted higher signature 5* activity among smokers, which may reflect increased activity of lower-fidelity repair pathways in the setting of increased levels of tobacco-mediated DNA damage.

Together, our data suggest that the genomic imprint of signature 5* depends on both the extent of DNA damage (from tobacco or other mutagens) and the relative activity of high- and low-fidelity DNA repair pathways, which is altered in the setting of an *ERCC2* mutation. Further studies will be needed to characterize the mechanisms underlying signature 5* activity in tumors that lack an *ERCC2* mutation and to explore potential relationships between signature 5* activity and clinically relevant endpoints such as treatment response.

URLs. COSMIC mutational signatures database, http://cancer.sanger.ac.uk/cosmic/signatures; Broad Institute TCGA Genome Data Analysis Center, http://firebrowse.org/; UniProt, http://www.uniprot.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

G.G. and J.K. were partially funded by the NIH TCGA Genome Data Analysis Center (U24CA143845). P.P. and A.K. were funded by the startup funds of G.G. at Massachusetts General Hospital. K.W.M. was partially funded by an American Society of Clinical Oncology (ASCO) Young Investigator Award and an American Society of Radiation Oncology (ASTRO) Junior Faculty Career Research Training Award. J.E.R. was partially funded by the Starr Cancer Consortium and the Memorial Sloan Kettering Geoffrey Beane Center. E.M.V.A. was partially funded by a Damon Runyon Clinical Investigator Award. A.D.D'A. was partially funded by the Starr Cancer Consortium. G.G. was partially funded by the Paul C. Zamecnik, MD, Chair in Oncology at Massachusetts General Hospital.

AUTHOR CONTRIBUTIONS

J.K. conceived the work, performed analyses, and wrote the manuscript. K.W.M. conceived the work, performed analyses, and wrote the manuscript. P.P. conceived the work, performed analyses, and wrote the manuscript. L.Z.B. performed analyses and edited the manuscript. A.K. performed analyses and edited the manuscript. G.T. performed analyses and edited the manuscript. D.J.K. contributed scientific insight and edited the manuscript. J.E.R. contributed scientific insight and edited the manuscript. E.M.V.A. conceived the work, contributed scientific insight, and edited the manuscript. A.D.D'A. conceived the work, contributed scientific insight, and edited the manuscript. G.G. conceived the work, oversaw the analyses, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
2. Dietlein, F., Thelen, L. & Reinhardt, H.C. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. *Trends Genet.* **30**, 326–339 (2014).
3. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
4. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
5. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
6. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
7. Marteijn, J.A., Lans, H., Vermeulen, W. & Hoeijmakers, J.H. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).
8. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
9. Fuss, J.O. & Tainer, J.A. XPB and XPD helicases in TFIIH orchestrate DNA duplex opening and damage verification to coordinate repair with transcription and cell cycle via CAK kinase. *DNA Repair (Amst.)* **10**, 697–713 (2011).

10. Compe, E. & Egly, J.M. TFIIH: when transcription met DNA repair. *Nat. Rev. Mol. Cell Biol.* **13**, 343–354 (2012).
11. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
12. Van Allen, E.M. *et al.* Somatic *ERCC2* mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. *Cancer Discov.* **4**, 1140–1153 (2014).
13. Guo, G. *et al.* Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nat. Genet.* **45**, 1459–1463 (2013).
14. Yap, K.L. *et al.* Whole-exome sequencing of muscle-invasive bladder cancer identifies recurrent mutations of *UNC5C* and prognostic importance of DNA repair gene mutations on survival. *Clin. Cancer Res.* **20**, 6605–6617 (2014).
15. Freedman, N.D., Silverman, D.T., Hollenbeck, A.R., Schatzkin, A. & Abnet, C.C. Association between smoking and risk of bladder cancer among men and women. *J. Am. Med. Assoc.* **306**, 737–745 (2011).
16. Ploeg, M., Aben, K.K. & Kiemeney, L.A. The present and future burden of urinary bladder cancer in the world. *World J. Urol.* **27**, 289–293 (2009).
17. Benhamou, S. *et al.* DNA adducts in normal bladder tissue and bladder cancer risk. *Mutagenesis* **18**, 445–448 (2003).
18. Lee, H.W. *et al.* Acrolein- and 4-aminobiphenyl-DNA adducts in human bladder mucosa and tumor tissue and their mutagenicity in human urothelial cells. *Oncotarget* **5**, 3526–3540 (2014).
19. Talaska, G., al-Juburi, A.Z. & Kadlubar, F.F. Smoking related carcinogen–DNA adducts in biopsy samples of human urinary bladder: identification of *N*-(deoxyguanosin-8-*yl*)-4-aminobiphenyl as a major adduct. *Proc. Natl. Acad. Sci. USA* **88**, 5350–5354 (1991).
20. Gao, W. *et al.* Genetic polymorphisms in the DNA repair genes *XPD* and *XRCC1*, *p53* gene mutations and bladder cancer risk. *Oncol. Rep.* **24**, 257–262 (2010).
21. Stern, M.C. *et al.* Polymorphisms in DNA repair genes, smoking, and bladder cancer risk: findings from the International Consortium of Bladder Cancer. *Cancer Res.* **69**, 6857–6864 (2009).
22. Tan, V.Y. & Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).
23. Nik-Zainal, S. *et al.* Association of a germline copy number polymorphism of *APOBEC3A* and *APOBEC3B* with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487–491 (2014).
24. Roberts, S.A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
25. Poon, S.L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
26. Schmeiser, H.H., Schoepe, K.B. & Wiessler, M. DNA adduct formation of aristolochic acid I and II *in vitro* and *in vivo*. *Carcinogenesis* **9**, 297–303 (1988).
27. Hoang, M.L. *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5**, 197ra102 (2013).
28. Poon, S.L. *et al.* Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.* **7**, 38 (2015).
29. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA* **112**, E5486–E5495 (2015).
30. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
31. Pfeifer, G.P. *et al.* Tobacco smoke carcinogens, DNA damage and *p53* mutations in smoking-associated cancers. *Oncogene* **21**, 7435–7451 (2002).
32. Francioli, L.C. *et al.* Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
33. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**, 514–517 (2003).
34. Haradhvala, N.J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
35. Polak, P. & Arndt, P.F. Transcription induces strand-specific mutations at the 5′ end of human genes. *Genome Res.* **18**, 1216–1223 (2008).
36. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
37. Groenendijk, F.H. *et al. ERBB2* mutations characterize a subgroup of muscle-invasive bladder cancers with excellent response to neoadjuvant chemotherapy. *Eur. Urol.* **69**, 384–388 (2016).
38. Plimack, E.R. *et al.* Defects in DNA repair genes predict response to neoadjuvant cisplatin-based chemotherapy in muscle-invasive bladder cancer. *Eur. Urol.* **68**, 959–967 (2015).
39. Bellmunt, J. *et al.* Gene expression of *ERCC1* as a novel prognostic marker in advanced bladder cancer patients receiving cisplatin-based chemotherapy. *Ann. Oncol.* **18**, 522–528 (2007).
40. Walsh, C.S. *et al. ERCC5* is a novel biomarker of ovarian cancer prognosis. *J. Clin. Oncol.* **26**, 2952–2958 (2008).
41. Jansen, J.G., Tsaalbi-Shtylik, A. & de Wind, N. Roles of mutagenic translesion synthesis in mammalian genome stability, health and disease. *DNA Repair (Amst.)* **29**, 56–64 (2015).
42. Sale, J.E., Lehmann, A.R. & Woodgate, R. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat. Rev. Mol. Cell Biol.* **13**, 141–152 (2012).

ONLINE METHODS

Data sets. Mutation data and relevant clinical data were downloaded from the Broad Institute TCGA Genome Data Analysis Center for the TCGA-130 cohort and from the journal websites for the DFCI/MSK-50 and BGI-99 cohorts and are summarized for all cases in **Supplementary Table 3** (refs. 11–13). We considered only coding mutations in mutation signature discovery and non-silent mutations in signature enrichment analysis.

Mutation signature analysis. *Methods and algorithms.* Mutational signature discovery is a process of deconvoluting cancer somatic mutation counts, stratified by mutation context or biologically meaningful subgroup, into a set of characteristic patterns (signatures) and inferring the activity of each of the discovered signatures across samples. Several groups, including ours, have used NMF to discover mutational processes[4,5,8,43]. We have recently described the use of a Bayesian version of NMF to discover mutational processes applied to chronic lymphocytic leukemia (CLL) data in Kasar *et al.*[5,22]. Below, we provide additional background and technical details regarding the Bayesian NMF methodology.

The common classification of SNVs is based on six base substitutions within the trinucleotide sequence context including the bases immediately 5′ and 3′ to the mutated base. Six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) each with 16 possible combinations of neighboring bases yield 96 possible mutation types (or contexts). Thus, the input data for mutation signature discovery comprise a $96 \times M$ matrix X, where M is the number of samples and each element x_{ij} represents the number of observed mutations of context i in sample j.

Because a collection of somatic mutations in a cancer genome is the outcome of multiple mutagenic processes operating over the lifetime of a patient, the mutation load x_{ij} is a superposition of signature-driven mutation burdens x_{ij}^k ($k = 1, 2, \ldots, K$) derived from K latent (unobserved) mutagenic processes. We further assume that signature-driven mutations x_{ij}^k are generated by a Poisson process parameterized by the context- and sample-specific rates $y_{ij}^k = w_{ik}h_{kj}$, where w_{ik} and h_{kj} denote the contribution of the kth mutagenic process to context i and its level of activity in sample j, respectively. Taken together, this model describes the observed mutations x_{ij} as the sum of the expected mutations y_{ij}^k as a consequence of K independent mutagenic operations plus background noise due to false positive mutation calls and other technical limitations. Accordingly, to detect the underlying mutational signatures, one needs to determine w_{ik} and h_{kj} for each signature as well as the unknown number of signatures, K. From the composite properties of the Poisson process, the distribution of total mutation load x_{ij} is also Poisson distributed with the total rate

$$y_{ij} = \sum_k w_{ik}h_{kj}$$

as $x_{ij} \sim \text{Poisson}(x_{ij}|y_{ij})$. Then, assuming that x_{ij} terms are independently conditioned on w_{ik} and h_{kj}, the log likelihood of the observed data $\boldsymbol{X}$, given the expectation $\boldsymbol{Y} = \boldsymbol{WH}$, factorizes and results in

$$\log(P(\boldsymbol{X}\,|\,\boldsymbol{Y})) = -D_{\mathrm{KL}}(\boldsymbol{X}\,|\,\boldsymbol{Y}) = -\sum_{ij} d(x_{ij}\,|\,y_{ij})$$

where $d(x|y)$ is the Kullback–Leibler (KL) divergence[22]. A maximum-likelihood approach for estimating $\boldsymbol{W}$ and $\boldsymbol{H}$ leads to an NMF problem of finding two non-negative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ that minimize the KL divergence between $\boldsymbol{X}$ and $\boldsymbol{WH}$, that is, $\min_{\boldsymbol{W},\boldsymbol{H}\geq 0} D_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{WH})$, where $\boldsymbol{W}$ and $\boldsymbol{H}$ correspond to the signature-loading and activity-loading matrices, respectively.

In the above formulation, the number of mutational processes or dimensionality K (also called the model 'complexity' or 'order') still remains unknown, and indeed the conventional NMF method requires K as an input[22]. A proper selection of K is important because using $K > K^{\mathrm{true}}$, where K^{true} is the true (unknown) underlying number of processes, will lead to overfitting, whereas accuracy will be impaired when using $K < K^{\mathrm{true}}$. To effectively address the issue of inferring the appropriate number of mutational signatures, we applied a Bayesian framework of NMF (Bayesian NMF) described by Tan and Fevotte to select an optimal K^* value that ensures the best explanation for the observed data $\boldsymbol{X}$ (ref. 22). Bayesian NMF exploits a 'shrinkage' or 'automatic relevance determination' technique to prune away irrelevant components in $\boldsymbol{W}$ and $\boldsymbol{H}$ that do not contribute to explaining $\boldsymbol{X}$. This pruning process is achieved by introducing relevance weights (or parameters), λ_k, each associated with the corresponding kth column in $\boldsymbol{W}$ and kth row in $\boldsymbol{H}$, and then imposing proper priors on $\boldsymbol{W}$, $\boldsymbol{H}$, and λ. During inference, columns and rows corresponding to irrelevant components rapidly shrink to zero as λ approaches its lower bound (which is close to zero and determined by the hyperparameters in the priors on λ), and the effective dimensionality K^* is automatically determined by the number of nonzero columns and rows in $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively[22].

The expected number of mutations associated with each mutational signature was determined after a scaling transformation, $\boldsymbol{X} \sim \boldsymbol{WH} = \tilde{\boldsymbol{W}}\tilde{\boldsymbol{H}}$, where $\tilde{\boldsymbol{W}} = \boldsymbol{WU}^{-1}$ and $\tilde{\boldsymbol{H}} = \boldsymbol{UH}$. The scaling matrix $\boldsymbol{U}$ is a $K \times K$ diagonal matrix with the element corresponding to the L_1-norm of column vectors of $\boldsymbol{W}$ (the sum of the elements of the vector). As a result, the kth column vector of the final signature matrix $\tilde{\boldsymbol{W}}$ represents a normalized profile of 96 trinucleotide mutation contexts associated with the kth signature (the profile vector sums to 1), and the kth row vector of the final activity matrix $\tilde{\boldsymbol{H}}$ represents the activity of the kth process across samples (the estimated, or expected, number of mutations generated by the kth process).

Moderating the effects of hypermutant samples on signatures. One of the challenges in discovering mutational signatures in a cohort of tumors with heterogeneous mutation burdens is the greater weight of hyper- or ultramutated samples in the discovered signatures. This greater weight can mask signals coming from samples with lower mutation burdens. To minimize this effect in our analysis, we applied a process moderating contributions from hypermutant samples to signature discovery, while preserving overall mutation counts in the cohort. More specifically, we first identified hyper- and ultramutated samples (outliers) as ones with

$$N_{\mathrm{SNV}} > N_{\mathrm{SNV}}^{\mathrm{median}} + 1.5 \times \mathrm{IQR}$$

where N_{SNV} is the number of SNVs in a given sample, $N_{\mathrm{SNV}}^{\mathrm{median}}$ is the median N_{SNV} across samples, and IQR represents the interquartile range. We then split mutation counts in each of the detected hypermutated samples into two separate columns with an equal number of mutations. This process was iterated, recalculating the median and IQR, until no hypermutated samples were detected, which resulted in the new mutation count matrix $\boldsymbol{X}^*$. It should be noted that this process preserves overall mutation counts across the cohort, while mutational loads in hyper- or ultramutated samples are equally partitioned into artificially created samples with the same spectra as their corresponding hypermutated samples. Because NMF is a linear dimensionality-reduction process, the original signature activity for hypermutated samples can be estimated by simply summing the activity of the artificially created samples derived from the original hypermutated sample.

Signature selection. We ran Bayesian NMF 50 times for the mutation count matrix $\boldsymbol{X}^*$, processed with the protocol in the above section with exponential priors for $\boldsymbol{W}$ and $\boldsymbol{H}$, and an inverse gamma prior for λ, starting from random initial conditions. The hyperparameter for the inverse gamma prior was set to $a = 10$, and the iterations were terminated when the tolerance for λ became less than 10^{-7}. All 50 runs in both the TCGA-130 and DFCI/MSK-50 cohorts converged to the solution with $K^* = 4$, and among the 50 solutions we selected for downstream analyses $\boldsymbol{W}$ and $\boldsymbol{H}$ that had the maximum posterior probability (**Fig. 1b** and **Supplementary Fig. 3b**)[22]. For the BGI-99 cohort, 44 of 50 runs converged to the solution with $K^* = 4$, while 6 runs converged to $K^* = 3$. After manually reviewing the signatures, we selected the maximum posterior solution with $K^* = 4$ (**Supplementary Fig. 5b**). We also performed mutational signature discovery separately for the combined cohort (COMB-279) and the combined cohort of muscle-invasive samples (COMB-MI-242) for signature comparison. In both cohorts, all 50 Bayesian NMF runs converged to the solution with $K^* = 4$. We also analyzed the combined cohort of TCGA-130 and DFCI/MSKCC-50 samples to investigate the association between smoking status and the activity of signature 5*, and here, as well, all 50 runs converged to the solution $K^* = 4$.

Signature enrichment analysis. The underlying correlation between the activity of a particular signature and the overall mutation burden can significantly confound the search for genes whose mutation status is associated

with the activity of the signature (**Fig. 2** and **Supplementary Figs. 2** and **9**). A straightforward statistical test that compares, for each gene, the distribution of signature activities between samples in which the gene is wild type versus mutant yields an inflation of significant *P* values for signatures that are correlated with overall mutation burden. This inflation is due to the fact that, in general, genes are more likely to be mutated in samples that have a higher mutation burden. To eliminate this inflation, we designed a permutation test in which we controlled both the gene-specific and sample-specific mutation counts when generating random permutations of the observed gene × sample binary mutation matrix, following an approach described in Strona *et al.*[44]. We used as a test statistic *T*, the one-tailed Wilcoxon rank-sum *P* value comparing the signature activities of mutant and wild-type samples of a given gene. We calculated this test statistic for the observed data, $T_{observed}$, as well as for every realization of the permuted mutation matrix, T^r_{random}, where $r = 1, 2, \ldots, 10^5$ (the total number of permutations). The final *P* value assigned to the gene was the fraction of permuted realizations with a test statistic equal to or more extreme than the observed test statistic (ones for which $T^r_{random} \leq T_{observed}$). By maintaining the row and column margins of the observed mutation matrix in every random realization, we corrected for the higher tendency of genes to be mutated in samples with higher mutation burden, as evidenced by the fact that nearly all genes except *ERCC2* are on the diagonals of the quantile–quantile plots in **Supplementary Figures 2** and **9**. Because of statistical power and computational efficiency considerations, we analyzed only genes with a non-silent mutation frequency of >5% across the analyzed cohort. We corrected for multiple-hypothesis testing using the Benjamini–Hochberg procedure and used FDR $q < 0.1$ as the significance threshold.

Our signature enrichment analysis identified *ERCC2* as the top significant gene associated with the activity of signature 5* across three independent cohorts (TCGA-130, DFCI/MSK-50, and BGI-99) and two combined cohorts (COMB-MI-242 and COMB-279) (**Fig. 3** and **Supplementary Figs. 4** and **8**). In fact, *ERCC2* was the only gene with FDR $q < 0.1$ across all five cohorts.

Once *ERCC2* was identified as the gene whose mutation status was most significantly associated with signature 5* activity, we used the Wilcoxon rank-sum test in assessing downstream associations between smoking status (in samples with mutant or wild-type *ERCC2*) and overall mutation burden (**Fig. 5a**) or signature 5* activity (**Fig. 5b**,**c** and **Supplementary Figs. 12** and **13**).

Clustering analysis. Comparison of the signatures discovered (**Supplementary Fig. 1**) in five cohorts (TCGA-130, DFCI/MSKCC-50, BGI-99, COMB-MI-242, and COMB-279) and 30 COSMIC signatures was performed using the standard hierarchical clustering R package with a distance of 'cosine' similarity and 'average' linkage options. The clustering analyses based on mutations attributed to signature 5* (**Supplementary Fig. 6a**,**b**) or the total number of SNVs (**Supplementary Fig. 6c**,**d**) across 96 mutation contexts were performed using a 'Euclidian' distance and 'ward.D' linkage method.

Structure modeling and CLUMPS analysis. As a basis for structural modeling of the ERCC2 protein, we used the crystal structure of the homologous protein XPD/Rad3-related DNA helicase (UniProt, Q4JC68) from *Sulfolobus acidocaldarius* (Protein Data Bank (PDB), 3CRV). *ERCC2* mutations were mapped to the bacterial protein on the basis of a global sequence alignment of the two proteins using the UniProt alignment tool with default parameters. To assess the significance of the spatial clustering of residues affected by missense mutations, we used the CLUMPS method[29]. Briefly, CLUMPS summarizes all pairwise Euclidean distances (transformed by a Gaussian function) between the centroids of mutated residues into a weighted average proximity (WAP) score and compares the score to a null model of random mutation scattering across all residues in the structure to calculate an empirical *P* value. In this study, we modified CLUMPS by using signature 5* activity instead of mutation recurrence levels to calculate the WAP score. The weight of each mutated residue *r* was calculated as $n_r = \text{Sig5}_r/\max(\text{Sig5})$, where Sig5_r is the signature 5* activity of the sample with the mutation *r* and max(Sig5) is the maximal value across all mutated residues. In cases where multiple samples had missense mutations affecting the same residue, the average Sig5_r value for these samples was used.

Forced deconvolution of signature 5* activity into COSMIC signature 4 and 5 contributions. Projection of the activity of signature 5* onto COSMIC signatures 4 and 5 was performed in the combined TCGA-130 and DFCI/MSK-50 cohort (the 180 cases with known smoking status). We used the NMF method[45] on the squared error divergence, with a fixed signature-loading matrix, $\boldsymbol{W}^*$ (96 × 2), where the column vectors corresponded to normalized COSMIC signatures 4 and 5. We used the estimated mutation counts of signature 5*, $\boldsymbol{X}_{5^*}$ (96 × 180), as an input matrix to NMF. Then, the activity-loading matrix $\boldsymbol{H}^*$ (2 × 180) was determined by standard NMF iteration of the multiplicative update algorithm, resulting in $\boldsymbol{X}_{5^*} \sim \boldsymbol{W}^*\boldsymbol{H}^*$. The row vectors in $\boldsymbol{H}^*$ represent the deconvolution of the activity of signature 5* onto COSMIC signatures 4 and 5.

Germline enrichment analysis. We identified all germline variants in 28 manually curated NER genes: *ERCC1*, *ERCC2*, *ERCC3*, *ERCC4*, *ERCC5*, *ERCC6*, *ERCC8*, *DDB1*, *DDB2*, *GTF2H1*, *GTF2H2*, *GTF2H3*, *GTF2H4*, *GTF2H5*, *LIG1*, *RAD23A*, *RAD23B*, *XPA*, *XPC*, *CETN2*, *CUL4B*, *CUL4A*, *CDK7*, *MNAT1*, *UVSSA*, *MMS19*, *ERCC6–PGBD3*, and *BIVM–ERCC5*. For this analysis, we considered only rare variants, defined as those present at <2% frequency in the TCGA-130 and DFCI/MSKCC-50 combined cohort (total of 180 samples). To identify overall enrichment of NER germline variants in samples with high signature 5* activity, we first computed the running enrichment score (ES) for somatic *ERCC2* mutations[46], which quantifies the degree to which somatic *ERCC2* mutations are over-represented in samples with high signature 5* activity (**Supplementary Fig. 11a**). The rank at the maximum running ES score, $R^* = 53$, was chosen to divide samples into groups containing samples with high signature activity (rank $\leq R^*$) and low signature activity (rank $> R^*$). The overall enrichment of NER pathway germline variants was assessed using a one-tailed Fisher's exact test with a 2 × 2 contingency table for *ERCC2* mutation status and sample group. We also repeated the same statistical test after removing samples with somatic *ERCC2* mutations to examine enrichment of NER germline variants in samples with wild-type *ERCC2*.

Because the functional effects of specific germline variants vary depending on the resulting amino acid changes, we performed a separate enrichment analysis by further stratifying the germline variants by resulting amino acid change. Variant-level signature 5* enrichment analysis was then performed for recurrent variants (≥2% frequency) by comparing the activity of signature 5* in samples with a specific germline variant and the remaining samples using a one-tailed Wilcoxon rank-sum test. To eliminate the contribution of *ERCC2* somatic mutations to signature enrichment, the analysis was restricted to samples with wild-type *ERCC2*, identifying several germline variants that were associated with signature 5* activity (**Supplementary Fig. 11b**).

Estimation of clonality using ABSOLUTE. Tumor samples are frequently contaminated with normal cells. ABSOLUTE infers the purity and ploidy of these heterogeneous populations using copy number and mutation data[47]. ABSOLUTE also estimates local copy number in cancer cells and the CCF of each mutation (the fraction of cancer cells harboring the mutation). To determine clonal versus subclonal mutation status for the 126 TCGA samples with available data, we followed the procedure described by Landau *et al.*[48] Specifically, mutations with Pr(CCF >0.95) >0.5 were annotated as clonal, whereas others were considered subclonal. The enrichment analysis of clonal signature 5* mutations in samples with clonal *ERCC2* mutations (**Fig. 6** and **Supplementary Fig. 17**) was performed by pairwise comparison of the number of clonal and subclonal mutations attributed to signature 5* in samples with clonal *ERCC2* mutations using the two-tailed pairwise Mann–Whitney test.

Multivariate regression analysis. Age, sex, smoking status, and *ERCC2* mutation status were considered as regression variables to explain the activity of signature 5* as a response variable in a multivariate linear regression model. Regression was performed using the standard R package.

doi:10.1038/ng.3557

Transcription strand bias analysis. We reran Bayesian NMF in the muscle-invasive combined cohort COMB-MI-242, further stratifying mutations by transcriptional strand (positive strand (+) or negative strand (−)), resulting in a total of 192 mutation contexts—96(+) and 96(−) contexts. Here the negative strand refers to the transcribed (template) strand, while the positive strand refers to the non-transcribed strand. For example, C>A(−) mutations at a GCT motif are added with G>T(+) mutations at an AGC motif, while C>A(+) mutations at a GCT motif are added with G>T(−) mutations at an AGC motif. Then, the transcription strand bias for C>A mutations at a GCT motif was defined as the ratio of the estimated number of C>A(−) mutations at a GCT motif to the estimated number of C>A(+) mutations at a GCT motif. As in the analysis with 96 contexts, all 50 Bayesian NMF runs with 192 contexts converged to a $K^* = 4$ solution (**Supplementary Fig. 15a**). The resulting signatures showed the strongest transcriptional strand bias for C>A and T>C mutations (**Supplementary Fig. 15b**).

Code availability. The basic source code for signature discovery will be available at the Broad Institute's Cancer Genome Analysis website, https://www.broadinstitute.org/cancer/cga/.

43. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
44. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).
45. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
46. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
47. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
48. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).