Somatic ERCC2 Mutations Are Associated with a Distinct Genomic Signature in Urothelial Tumors

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2016 October 25.

Published in final edited form as:

Nat Genet. 2016 June ; 48(6): 600–606. doi:10.1038/ng.3557.

Somatic *ERCC2* Mutations Are Associated with a Distinct Genomic Signature in Urothelial Tumors

Jaegil Kim[#1], **Kent W Mouw**[#2,3], **Paz Polak**[#1,3,4,5], **Lior Z Braunstein**[1,3], **Atanas Kamburov**[1,3,4,5], **David J Kwiatkowski**[3,6], **Jonathan E Rosenberg**[7], **Eliezer M Van Allen**[1,3,8], **Alan D'Andrea**[2,3,9], and **Gad Getz**[1,3,4,5,11]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2]Department of Radiation Oncology, Brigham & Women's Hospital, Dana-Farber Cancer Institute, Boston, MA, USA

[3]Harvard Medical School, Boston, MA, USA

[4]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

[5]Cancer Center, Massachusetts General Hospital, Boston, MA, USA

[6]Division of Pulmonary Medicine, Brigham & Women's Hospital, Boston, MA, USA

[7]Genitourinary Oncology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[9]Center for DNA Damage and Repair, Dana-Farber Cancer Institute, Boston, MA, USA

[#] These authors contributed equally to this work.

Abstract

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 reveals a

[11]Corresponding Author: Gad Getz, PhD, Cancer Program, Broad Institute of MIT and Harvard, 415 Main St., Cambridge, MA 02142, gadgetz@broadinstitute.org.

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. D.J.K. contributed scientific insight and edited the manuscript. J.E.R. contributed scientific insight and edited the manuscript. E.V.M. conceived the work, contributed scientific insight, and edited the manuscript. A.D. 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 E.V.M. and J.E.R. have ownership interest in a patent (pending) for use of *ERCC2* mutational status as a clinical biomarker.

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

strong association between somatic *ERCC2* mutations and activity of a mutational signature characterized by a broad spectrum of base changes. In addition, we note an association between 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 the first NER-related mutational signature and highlight the related roles of DNA damage and subsequent DNA repair in shaping the tumor mutational landscape.

Introduction

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 the 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 nucleotide excision repair (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 common 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 *ERCC2*, a DNA helicase that plays a central role in the NER pathway by unwinding the DNA duplex adjacent to the 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 US and are associated with exposure to tobacco, chemicals, and certain infectious agents[15, 16]. Many of these carcinogens are known to damage DNA through 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 wildtype (WT) *ERCC2* but 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 been previously described. Unbiased enrichment analysis identified *ERCC2* as the gene which, 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, the first description of a tobacco-related mutational signature in urothelial cancer. Together, these findings identify the first NER-related mutational signature and underscore the importance of both exposure to DNA damaging agents and operation of DNA repair pathways on 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 Cancer Genome Atlas (TCGA-130)(Figure 1a, Supplementary Table 1). We applied a Bayesian variant of the NMF algorithm to mutation counts, stratified by 96 tri-nucleotide mutational contexts, in order 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 (i.e., the estimated number of mutations associated with each signature) (**Methods**).[5, 22]

Our analysis identified four independent mutational signatures in the TCGA-130 cohort (Figure 1b; Supplementary Tables 2, 3) and although our analysis methods are not identical, the signatures match four signatures previously identified by the Sanger Institute (cosine similarities between 0.86 and 0.95) and described in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Supplementary Figure 1; 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 Figure 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-methyl-cytosine deamination events (C>T CpG in Figure 1b; COSMIC signature 1). Finally, a fourth signature was identified which closely resembles COSMIC signature 5 (cosine similarity 0.90; denoted as signature 5* in Figure 1b and Supplementary Figure 1). COSMIC signature 5 is characterized by a broad spectrum of base changes and is present in all tumor types, but an etiology has not yet been described.

Signature 5* Activity Is Associated with ERCC2 Mutations

In order to further characterize signature 5*, we performed signature enrichment analysis to identify genes which, when mutated, were associated with increased activity of signature 5* (**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 which

carried a non-silent mutation in the gene versus those that did not. To ensure that the increased signature 5* activity did not reflect an increase in overall mutation burden, we assessed the significance level using a permutation-based method that maintains the overall number of non-silent mutations per sample and gene (**Methods**). *ERCC2* was the only significant gene (Benjamini-Hochberg False Discovery Rate Q=8.6×10^{-3}, P=3×10^{-5}; Figure 2; Supplementary Figure 2). Overall, 16 of 130 tumors had a non-silent *ERCC2* mutation, and these tumors had an increase of 95 mutations (135 vs 40) in the median activity of signature 5* (Figure 3a).

Validation in Two Independent Cohorts of Urothelial Tumors

In order 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 is comprised of 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 0.93–0.99; Supplementary Figures 1, 3; Supplementary Table 5). Repeating the gene mutation enrichment analysis with signature 5* activity identified three significant genes (Q≤0.1), with *ERCC2* being the most significant (Q=0.042, P=1.9×10^{-4}; Figure 2; Supplementary Figures 2, 4). Nine of 50 tumors had a non-silent mutation in *ERCC2* and an increase of 188 mutations (220 vs 32) in median signature 5* activity.

The second validation cohort was comprised of 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 Figure 5). The first two resembled the two APOBEC-associated signatures (cosine similarities 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 0.96) and has been linked to exposure to aristolochic acid (AA), an ingredient in some food supplements that are most commonly used in Asian countries.[25] Indeed, consumption of AA has been associated with increased risk of urothelial cancers.[26–28] The fourth signature in this cohort appears to be a superposition of the two other signatures identified in the previous cohorts, C>T CpG and signature 5*, with 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* (10 tumors with a non-silent *ERCC2* mutation and a median increase of 55 mutations per sample; Q=0.012, P=2.5×10^{-4}; Figure 2; Supplementary Figures 2, 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 compared to WT *ERCC2* tumors (124 vs 33), providing the strongest statistical evidence for the association between *ERCC2* and signature

5* activity (Q=1.6×10^{-3}, P=1.0×10^{-5}; Fig 2; Supplementary Figure 2, 4). Together, these data strongly suggest that although signature 5* activity is present in both WT and mutant *ERCC2* tumors, 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 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=1.7×10^{-12}, Fisher's exact test; Supplementary Figure 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=4.4×10^{-14}, Supplementary Fig 6b). Although *ERCC2* mutations are associated with higher overall mutation burden[11–13], segregation was not driven by it, as *ERCC2* mutated tumors segregated less strongly when clustering was performed using the total number of SNVs ($P_{COMB\text{-}279}$=0.1, $P_{COMB\text{-}MI\text{-}242}$=0.008; Supplementary Figure 6c, d).

All but one of the 35 non-silent *ERCC2* mutations across cohorts were missense mutations, and most (25 of 34) were located within or adjacent to (±10 amino acids) the conserved helicase motifs, suggesting that the mutations may have an impact on *ERCC2* protein function (Supplementary Figure 7a). Supporting this hypothesis, helicase motif mutations were associated with higher signature 5* activities compared to mutations located elsewhere in the protein (median no. signature 5* mutations 134 vs 96, p=0.037). To assess the spatial relationship of the mutations, we utilized CLUMPS, a novel algorithm for assessing spatial clustering of mutations within 3D protein structures, and found that the *ERCC2* mutations were significantly clustered (p=0.0026), further suggesting a functional role (Supplementary Figure 7b; **Methods**).[29]

For each of the three cohorts analyzed here, *ERCC2* mutated tumors have been shown to have an increase in overall mutation burden compared to WT *ERCC2* tumors.[11–13] We asked if this increase was due solely to increased signature 5* activity or whether activities of other signatures were also increased. Indeed, activity of the APOBEC2 signature was also increased in *ERCC2* mutated tumors compared to WT *ERCC2* tumors in the TCGA-130 cohort (39 vs 12, *P*=0.004; Figure 3b); however, unlike the association between *ERCC2* and signature 5*, the association of *ERCC2* with the APOBEC2 signature was not significant after correcting for multiple testing (Q=0.54). A similar association between *ERCC2* 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 Figures 8, 9). There was no increase in activity of the APOBEC1 (78 vs 103 in TCGA-130, *P*=0.99) or C>T CpG (0 vs 13, P=0.91) signatures in mutant vs WT *ERCC2* tumors in any of the cohorts (Fig 3b; Supplementary Figure 8). These results demonstrate that the increase in 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 may 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* and mutations in any individual NER gene or the pathway as a whole (when *ERCC2* is excluded)(Figure 4; Supplementary Figure 10). However, anecdotally, among the 20 WT *ERCC2* tumors with the highest signature 5* activity, six had a mutation in a different gene in the NER pathway. In addition, germline data was available for the TCGA-130 and DFCI/MSK-50 cohorts, and there was a trend towards an association between rare (<2% frequency in the cohorts) NER germline variants and signature 5* activity in WT *ERCC2* cases (19 of the 32 WT *ERCC2* tumors with highest signature 5* activity had a NER germline variant versus only 54 of the remaining 123 WT *ERCC2* tumors, p=0.086; **Methods;** Supplementary Figure 11a). Moreover, four specific NER germline alleles were enriched (Q<0.1) in WT *ERCC2* tumors with high signature 5* activity, and three of the four are predicted to be functionally deleterious (Supplementary Figure 11b).[30] However, additional studies in larger cohorts will be needed to further characterize the 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. Smoking status was available for the TCGA-130 and DFCI/MSK-50 cohorts, and they were therefore analyzed together. There was no difference in overall mutation burden in tumors from patients with any smoking history (`smokers') versus no smoking history (`non-smokers')(*P*=0.27, Wilcoxon rank-sum test; Figure 5a). However, activity of signature 5* was significantly higher in smokers than non-smokers (median number of signature 5* mutations: 49 vs 33, p= 0.009; Figure 5b), although the effect size is modest compared to that of an *ERCC2* mutation (Figure 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 non-smokers (Supplementary Figure 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 in any of the urothelial cohorts analyzed here.[4, 31] Given the association of signature 5* activity with smoking, we explored whether signature 5* includes a contribution from COSMIC signature 4 and that these processes were not separated by NMF due to insufficient power. To test this hypothesis, we attempted to separate signature 5* mutations into contributions from COSMIC signatures 4 and 5 (**Methods**). This analysis confirmed the close similarity of signature 5* to COSMIC signature 5 and revealed 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 Figure 14).

Several mutational signatures exhibit an asymmetric pattern of mutations on the transcribed versus non-transcribed DNA strand, a phenomenon that is attributed to increased high-fidelity repair of the transcribed strand by the transcription-coupled repair (TCR) subpathway of NER.[7, 32–35] To determine if signature 5* exhibits transcriptional strand bias, we repeated the Bayesian NMF analysis using 192 mutational contexts (instead of 96) by considering mutations on the transcribed and non-transcribed strands independently (**Methods**; Supplementary Figure 15). Signature 5* exhibits strand asymmetry in several contexts, including a bias for T>C transitions on the transcribed strand, as described for COSMIC signature 5.[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 decreased 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 as well as the length of time over which it operates. Recently, activity of COSMIC signatures 1 and 5 were found to be correlated with patient age, suggesting that the underlying mutational processes are active across the lifetime of somatic cells.[36] However, an association between age and COSMIC signature 5 activity was not 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 reveal an association between age and signature 5* activity (P=0.65, Supplementary Figure 16). Similarly, on multivariate regression analysis, *ERCC2* mutational status (P=3.5×10^{-14}) and smoking (P=0.038) were significantly associated with signature 5* activity, while age (P=0.60) and gender (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 (**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 probability[CCF≥0.95]>0.5) and five 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 WT *ERCC2* (clonal:subclonal ratio~1.9, P=0.49; Figure 6, Supplementary Figure 17). Overall, these data suggest that somatic *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 increased signature 5* activity in the 25 cisplatin responders compared to the 25 non-responders (P=0.027; Supplementary Figure 18); however, signature 5* activity was not associated with cisplatin response in WT *ERCC2* cases (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 datasets, hence called `signature 5*' here) 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, signature 5 activity has been identified in all tumor types characterized to date. 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 (TLS) that normally operate in parallel with high-fidelity repair pathways like NER, and are upregulated when high-fidelity repair is compromised.[41, 42] In urothelial cancer, somatic *ERCC2* mutations appear 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 appear 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 was modest relative to the effect of 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 activity of a specific mutational signature in urothelial tumors had not been previously described. Here, we noted an increase in 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 others mutagens) as well as 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.

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.[11–13] We considered only coding mutations in the mutation signature discovery and non-silent mutations in signature enrichment analysis.

Mutation Signature Analysis

(1) Methods and Algorithms—The mutational signatures discovery is a process of deconvoluting cancer somatic mutations counts, stratified by mutation contexts or biologically meaningful subgroups, 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 non-negative matrix factorization (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 tri-nucleotide sequence context including the base immediately 5' and 3' to the mutated base. Six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) with 16 possible combinations of neighboring bases result in 96 possible mutation types (or contexts). Thus the input data for the mutation signature discovery is a 96-by-M mutation matrix $\boldsymbol{X}$, where M is the number of samples, and each element x_{ij} represents the number of observed mutations of context i in sample j.

Since a collection of somatic mutations in a cancer genome is an 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 the latent (unobserved) K mutagenic processes. We further assume that signature-driven mutations x_{ij}^{k} are generated by a Poisson process parameterized by a product of context- and sample-specific rates $y_{ij}^{k}=w_{ik}h_{kj}$, where w_{ik} and h_{kj} denote the contribution of k-th

mutagenic process on context *i* and its level of activity in sample *j*. Taken together, this model describes the observed mutations x_{ij} by the sum of the expected mutations y_{ij}^{k} as a consequence of independent *K* mutagenic operations plus a background noise due to false positive mutation calls and other technical limitations. Accordingly, in order 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

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

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

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

where $d(x|y)$ is the Kullback-Leibler (KL) divergence.[22] A maximum likelihood approach for estimating ***W*** and ***H*** leads to a non-negative matrix factorization (NMF) problem to find two non-negative matrices ***W*** and ***H*** that minimize the KL divergence between ***X*** and ***WH*** - i.e., $\min_{\boldsymbol{W,H}>=\boldsymbol{0}} D_{KL}(\boldsymbol{X}|\boldsymbol{WH})$ where ***W*** and ***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*), remains still unknown, and indeed the conventional NMF method requires *K* as an input.[22] A proper selection of *K* is important since using $K > K^{\text{true}}$, where K^{true} is the true (unknown) underlying number of processes, will lead to overfitting, while accuracy will be impaired when using $K < K^{\text{true}}$. In order 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** that ensures the best explanation for the observed data ***X***.[22] Bayesian NMF exploits a *shrinkage* or *automatic relevance determination* (ARD) technique to prune away irrelevant components in ***W*** and ***H*** that do not contribute to explaining ***X***. This pruning process is achieved by introducing relevance weights (or parameters), λ_k, each associated with the corresponding *k*-th column in ***W*** and *k*-th row in ***H***, and then imposing proper priors on ***W***, ***H***, and λ. During the 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 hyper-parameters in the priors on λ) and the effective dimensionality *K** is automatically determined by the number of non-zero columns and rows in ***W*** and ***H***, respectively.[22]

The expected number of mutations associated with each mutational signature was determined after a scaling transformation, $\boldsymbol{X} \tilde{} \boldsymbol{WH} = \tilde{W}\tilde{H}$ where $\tilde{W} = \mathbf{W}\mathbf{U}^{-1}$ and $\tilde{H} = \boldsymbol{UH}$. The scaling matrix ***U*** is a $K \times K$ diagonal matrix with the element corresponding to the L_1-norm of column vectors of ***W*** (ie. the sum of the elements of the vector). As a result, the *k*-th

column vector of the final signature matrix $\tilde{W}$ represents a normalized profile of 96 tri-nucleotide mutation contexts associated with the *k*-th signature (the profile vector sums to 1), and the *k*-th row vector of the final activity matrix $\tilde{H}$ represents the activity of the *k*-th process across samples (i.e., the estimated (or expected) number of mutations generated by the *k*-th process).

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

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

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

(3) Signature selection—We ran Bayesian NMF 50 times for the mutation count matrix ***X**** processed by the protocol in (2) with exponential priors for ***W*** and ***H***, and inverse gamma prior for the λ starting from random initial conditions. The hyper-parameter 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 TCGA-130 and DFCI/MSK-50 cohorts converged to the solution with K^*= 4 and among the 50 solutions we selected, for downstream analyses, the ***W*** and ***H*** that had the maximum posterior probability (Figure 1b and Supplementary Figure 3b).[22] For BGI-99 cohort, 44/50 runs converged to the solution with K^*= 4, while 6 runs converged to K^*= 3. After manually reviewing signatures, we selected the maximum posterior solution with K^*= 4 (Supplementary Figure 5b). We also separately performed the mutational signature discovery 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 (Figure 2 and Supplementary Figures 2, 9). A straightforward statistical test that compares, for each gene, the distribution of signature activities between samples in which the gene is wildtype 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 control both the gene-specific and sample-specific mutation counts when generating the random permutations of the observed gene-by-sample binary mutation matrix, following an approach described in Strona et al.[44] We use as a test statistic, T, the one-tailed Wilcoxon rank-sum p-value between the signature activities of mutant and wildtype samples of a given gene. We calculate 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, …,10^5 (the total number of permutations). The final p-value assigned to the gene is the fraction of permuted realizations with an equal or more extreme value of the test statistic (i.e., ones for which $T^r_{random} \leq T_{observed}$). Since we maintain row and column margins of the observed mutation matrix in every random realization, we correct 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 diagonal of the Q-Q plots in Supplementary Figures 2 and 9. Due to statistical power and computational efficiency considerations, we analyzed only genes with >5% non-silent mutation frequencies across the analyzed cohort. We corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure and used a False Discovery Rate (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) (Figure 3, Supp. Figures 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 with mutation status most significantly associated with signature 5* activity, we utilized the Wilcoxon rank-sum test in assessing downstream associations between smoking status (in *ERCC2* mutant or wildtype samples) and overall mutation burden (Figure 5a) or signature 5* activity (Figures 5b, 5c; Supplementary Figures 12, 13).

Clustering Analysis

Comparison of signatures discovered (Supplementary Figure 1) in five cohorts (TCGA-130, DFCI/MSKCC-50, BGI-99, COMB-MI-242, COMB-279) and 30 COSMIC signatures was performed using the standard hierarchical clustering R-package with a distance of “cosine”

similarity and the “average” linkage options. The clustering analyses based on mutations attributed to signature 5* (Supplementary Figure 6a, b) or the total number of single nucleotide variants (SNVs, Supplementary Figure 6c, d) across 96 mutation contexts was 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 (PDB: 3CRV). *ERCC2* mutations were mapped to the bacterial protein based on a global sequence alignment of the two proteins using the Uniprot alignment tool with default parameters. To assess the significance of spatial clustering of 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 $n_r = Sig5_r / max(Sig5)$, 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 in the same residue, the average $Sig5_r$ value over these samples was used.

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

Germline Enrichment Analysis

We identified all germline variants in 28 manually curated nucleotide excision repair (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 combined cohort of TCGA-130 and DFCI/MSKCC-50 (total 180 samples). To identify an overall enrichment of NER germline variants in samples with a 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 Figure 11a). The rank at the maximum running ES score, R*=53, was chosen to divide samples into the

signature-high (rank ≤ R*) and the signature-low (rank > R*) groups. The overall enrichment of NER pathway germline variants was assessed using a one-tailed Fisher's exact test with 2 × 2 contingency table for *ERCC2* mutation status and the sample grouping. We also repeated the same statistical test after removing samples with somatic *ERCC2* mutations in order to examine enrichment of NER germline variants in WT *ERCC2* samples.

Since the functional effects of specific germline variants vary depending on the resulting amino acid change, we performed a separate enrichment analysis by further stratifying the germline variants by the resulting amino acid change. The variant-level signature 5* enrichment analysis was then performed for recurrent variants (i.e., frequency ≥ 2) by comparing the activity of signature 5* between samples with a specific germline variant versus the remaining samples using a one-tailed Wilcoxon rank-sum test. To eliminate the contribution of *ERCC2* somatic mutations on the signature enrichment, the analysis was restricted to WT *ERCC2* samples, which identified several germline variants that were associated with signature 5* activity (Supplementary Figure 11b).

Estimation of Clonality using ABSOLUTE

Tumor samples are frequently contaminated with normal cells. ABSOLUTE infers the purity and ploidy of this heterogeneous population using copy number and mutation data.[47] ABSOLUTE also estimates local copy-number in the cancer cells and the cancer cell fraction (CCF) of each mutation (i.e., 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 probability(CCF>0.95) > 0.5 were annotated as clonal, while others were considered subclonal. The enrichment analysis of clonal signature 5* mutations in samples with clonal *ERCC2* mutations (Figure 6 and Supplementary Figure 17) was performed by pair-wise comparisons of the number of clonal versus subclonal mutations attributed to signature 5* in samples with clonal *ERCC2* mutations using the two-tailed pairwise Mann-Whitney test.

Multivariate Regression Analysis

The age, gender, 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. The regression was performed using the standard R-package.

Transcription Strand Bias Analysis

We re-ran the Bayesian NMF in the muscle-invasive combined cohort COMB-MI-242, but further stratified the mutations by their transcriptional strands (positive strand [+] or negative strand [−]), resulting in a total of 192 mutation contexts -- 96(+) and 96(−) contexts. Here, the negative strand (−) refers to transcribed (template) strand while the positive strand (+) refers to the non-transcribed strand. For example, C>A(−) mutations at GCT motif are added with G>T(+) mutations at AGC motif, while C>A(+) mutations at GCT motif are added with G>T(−) mutations at AGC motif. The transcription strand bias of C>A at GCT motif was defined as the ratio of the estimated number of mutations of C>A(−) at GCT to the estimated number of mutations of C>A(+) at GCT. As in the 96 context analysis, all 50 Bayesian NMF runs with 192 contexts converged to a K^*=4 solution (Supplementary Figure 15a). The

resulting signatures showed the strongest transcriptional strand bias in C>A and T>C mutations (Supplementary Figure15b).

Code Availability

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

G.G. and J.K. were partially funded by NIH TCGA Genome Data Analysis Center (U24CA143845). P.P. and A.K. were funded by G.G.'s startup funds at MGH. K.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.V.M. was partially funded by a Damon Runyon Clinical Investigator Award. G.G. was partially funded by the Paul C. Zamecnik, MD, Chair in Oncology at Massachusetts General Hospital.

References

1. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–15. [PubMed: 21720365]
2. Dietlein F, Thelen L, Reinhardt HC. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. Trends Genet. 2014; 30:326–39. [PubMed: 25017190]
3. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013; 153:17–37. [PubMed: 23540688]
4. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. [PubMed: 23945592]
5. Kasar S, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. 2015 in press.
6. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. 2014; 15:585–98. [PubMed: 24981601]
7. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat. Rev. Mol. Cell. Biol. 2014; 15:465–81. [PubMed: 24954209]
8. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–8. [PubMed: 23770567]
9. Fuss JO, Tainer JA. 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.). 2011; 10:697–713. [PubMed: 21571596]
10. Compe E, Egly JM. TFIIH: when transcription met DNA repair. Nat Rev Mol Cell Biol. 2012; 13:343–54. [PubMed: 22572993]
11. Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014; 507:315–22. [PubMed: 24476821]
12. Van Allen EM, et al. Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. Cancer Discov. 2014; 4:1140–53. [PubMed: 25096233]
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. 2013; 45:1459–63. [PubMed: 24121792]

14. Yap KL, 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. 2014; 20:6605–17. [PubMed: 25316812]
15. Freedman ND, Silverman DT, Hollenbeck AR, Schatzkin A, Abnet CC. Association between smoking and risk of bladder cancer among men and women. JAMA. 2011; 306:737–45. [PubMed: 21846855]
16. Ploeg M, Aben KK, Kiemeney LA. The present and future burden of urinary bladder cancer in the world. World J. Urol. 2009; 27:289–93. [PubMed: 19219610]
17. Benhamou S, et al. DNA adducts in normal bladder tissue and bladder cancer risk. Mutagenesis. 2003; 18:445–8. [PubMed: 12960413]
18. Lee HW, et al. Acrolein- and 4-Aminobiphenyl-DNA adducts in human bladder mucosa and tumor tissue and their mutagenicity in human urothelial cells. Oncotarget. 2014; 5:3526–40. [PubMed: 24939871]
19. Talaska G, al-Juburi AZ, Kadlubar FF. 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. 1991; 88:5350–4. [PubMed: 2052611]
20. Gao W, et al. Genetic polymorphisms in the DNA repair genes XPD and XRCC1, p53 gene mutations and bladder cancer risk. Oncol. Rep. 2010; 24:257–62. [PubMed: 20514470]
21. Stern MC, et al. Polymorphisms in DNA repair genes, smoking, and bladder cancer risk: findings from the international consortium of bladder cancer. Cancer Res. 2009; 69:6857–64. [PubMed: 19706757]
22. Tan VY, Fevotte C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. IEEE Trans. Pattern Anal. Mach. Intell. 2013; 35:1592–605. [PubMed: 23681989]
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. 2014; 46:487–91. [PubMed: 24728294]
24. Roberts SA, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45:970–6. [PubMed: 23852170]
25. Poon SL, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci. Transl. Med. 2013; 5:197ra101.
26. Schmeiser HH, Schoepe KB, Wiessler M. DNA adduct formation of aristolochic acid I and II in vitro and in vivo. Carcinogenesis. 1988; 9:297–303. [PubMed: 3338114]
27. Hoang ML, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci. Transl. Med. 2013; 5:197ra102.
28. Poon SL, et al. Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med. 2015; 7:38. [PubMed: 26015808]
29. Kamburov A, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. USA. 2015; 112:E5486–95. [PubMed: 26392535]
30. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 2009; 4:1073–81. [PubMed: 19561590]
31. Pfeifer GP, et al. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002; 21:7435–51. [PubMed: 12379884]
32. Francioli LC, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. 2015; 47:822–6. [PubMed: 25985141]
33. Green P, et al. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 2003; 33:514–7. [PubMed: 12612582]
34. Haradhvala NJ, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016; 164:538–49. [PubMed: 26806129]
35. Polak P, Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes. Genome Res. 2008; 18:1216–23. [PubMed: 18463301]
36. Alexandrov LB, et al. Clock-like mutational processes in human somatic cells. Nat. Genet. 2015

37. Groenendijk FH, et al. ERBB2 Mutations Characterize a Subgroup of Muscle-invasive Bladder Cancers with Excellent Response to Neoadjuvant Chemotherapy. Eur. Urol. 2015
38. Plimack ER, et al. Defects in DNA Repair Genes Predict Response to Neoadjuvant Cisplatin-based Chemotherapy in Muscle-invasive Bladder Cancer. Eur. Urol. 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. 2007; 18:522–8. [PubMed: 17229776]
40. Walsh CS, et al. ERCC5 is a novel biomarker of ovarian cancer prognosis. J. Clin. Oncol. 2008; 26:2952–8. [PubMed: 18565881]
41. Jansen JG, Tsaalbi-Shtylik A, de Wind N. Roles of mutagenic translesion synthesis in mammalian genome stability, health and disease. DNA Repair (Amst.). 2015; 29:56–64. [PubMed: 25655219]
42. Sale JE, Lehmann AR, Woodgate R. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. Nat. Rev. Mol. Cell. Biol. 2012; 13:141–52. [PubMed: 22358330]
43. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149:979–93. [PubMed: 22608084]
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. 2014; 5:4114. [PubMed: 24916345]
45. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401:788–91. [PubMed: 10548103]
46. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA. 2005; 102:15545–50. [PubMed: 16199517]
47. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 2012; 30:413–21. [PubMed: 22544022]
48. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013; 152:714–26. [PubMed: 23415222]

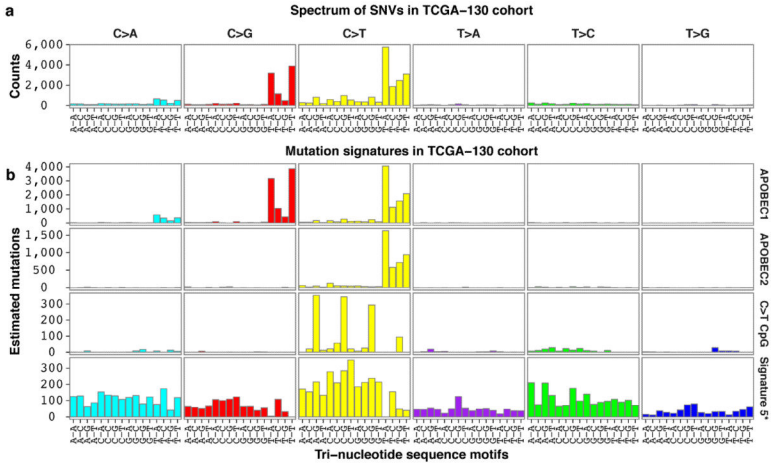

Figure 1.

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 as the mutated pyrimidine and the adjacent 3' and 5' bases. (**b**) A Bayesian non-negative matrix factorization algorithm was applied to identify signatures from the matrix of mutation counts across tumors. Four distinct mutational signatures were identified.

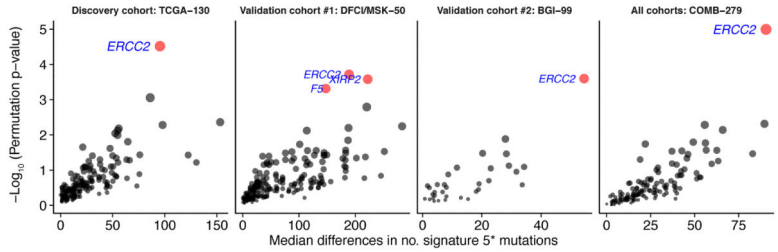

Figure 2.

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 a wild-type versus mutated copy of the gene while controlling for overall mutation burden per gene and sample. 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).

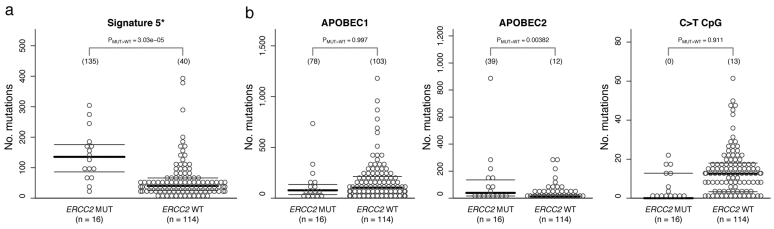

Figure 3.

Comparison of signature activities in wild-type (WT) versus mutant *ERCC2* tumors in the TCGA-130 cohort. (**a**) The estimated number of signature 5* mutations was significantly higher in *ERCC2* mutated tumors compared to WT *ERCC2* tumors. (**b**) Estimated number of mutations attributed to the other three 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.

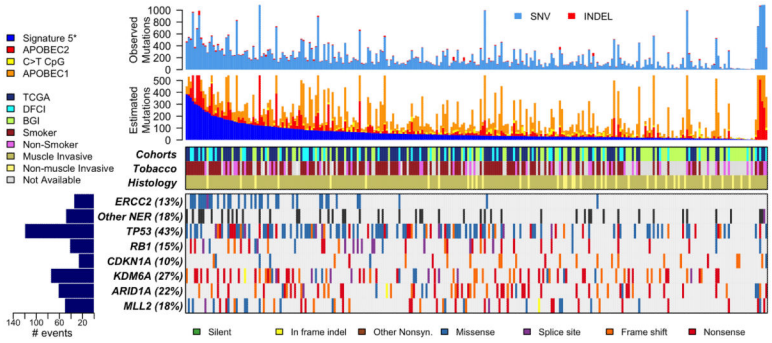

Figure 4.

Overall mutation rate, 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 arranged in descending order of signature 5* activity), cohort, smoking status, and stage (muscle invasive versus non-muscle invasive). In the bottom half of the figure, the mutational status of *ERCC2* and other genes of interest are color-coded by type of mutation. Somatic events in non-*ERCC2* NER pathway genes are collapsed in a single track (see Supplementary Figure 10 for expanded NER pathway gene list) and are followed by other significantly mutated genes in urothelial cancer (*TP53*, *RB1*, etc).

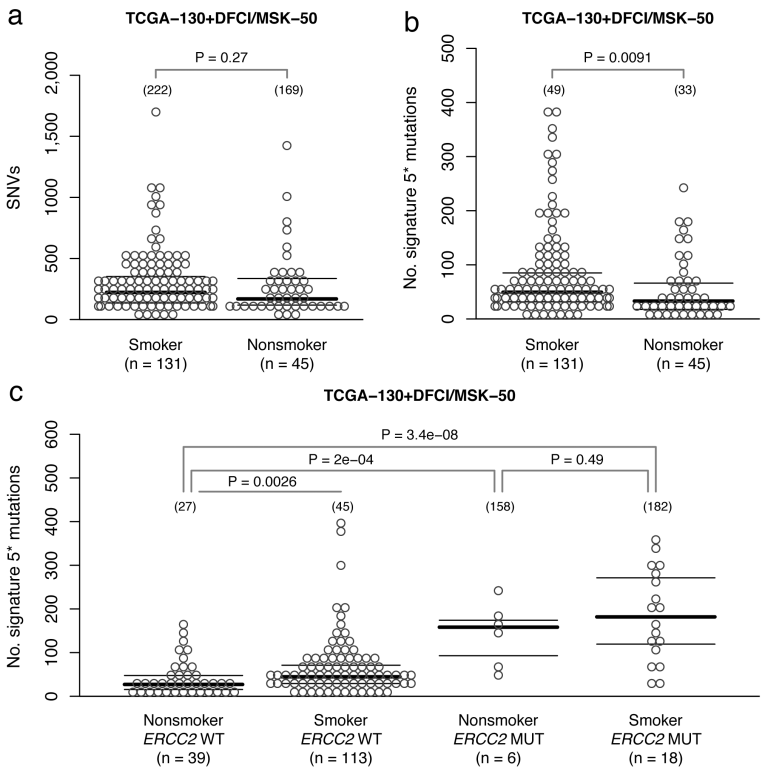

Figure 5.

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 compared to non-smokers in the combined TCGA-130 + DFCI/MSK-50 cohort. The median number of mutations is shown in parentheses and p-values were calculated using the Wilcoxon rank-sum test. (**b**) The estimated number of signature 5* mutations was significantly higher in smokers than in non-smokers. (**c**) Among patients with wild-type (WT) *ERCC2* tumors, the number of signature 5* mutations was significantly higher in smokers than non-smokers, 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* and signature 5*.

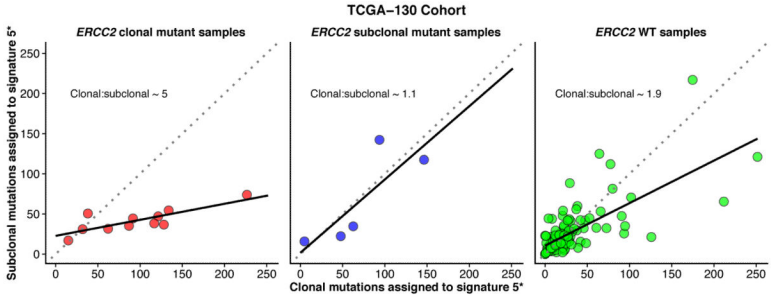

Figure 6.

Association between clonality of *ERCC2* mutations and clonality of signature 5* mutations. For tumors with a clonal *ERCC2* mutation (defined as probability[cancer cell fraction≥0.95]>0.5); red circles, left panel), the majority of signature 5* mutations were clonal (clonal:subclonal ratio~5). For tumors with a subclonal *ERCC2* mutation (blue circles, center panel) or WT *ERCC2* (green circles, right panel), the ratio of clonal to subclonal signature 5* mutations was much lower (clonal:subclonal ratio~1.1 and ~1.9, respectively).