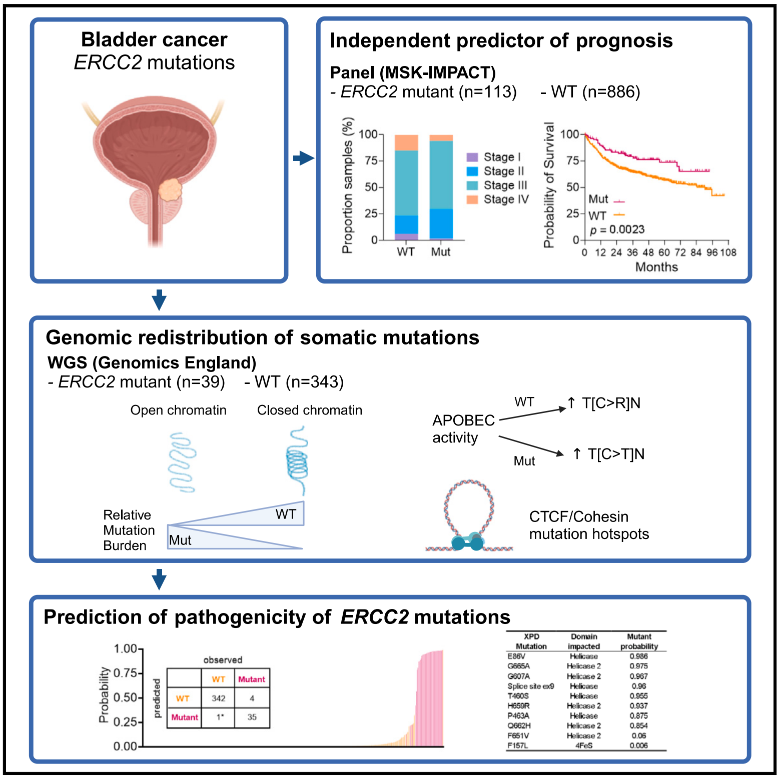
Cell Genomics

Article

ERCC2 mutations alter the genomic distribution pattern of somatic mutations and are independently prognostic in bladder cancer

Graphical abstract



	WT	Mutant
WT	342	4
Mutant	1*	35

XPD Mutation	Domain impacted	Mutant probability
E86V	Helicase	0.986
G665A	Helicase 2	0.975
G607A	Helicase 2	0.967
Splice site ex9	Helicase	0.96
T460S	Helicase	0.955
H659R	Helicase 2	0.937
P463A	Helicase	0.875
Q662H	Helicase 2	0.854
F651V	Helicase 2	0.06
F157L	4FeS	0.006

Highlights

- *ERCC2* mutations are an independent predictor of prognosis in bladder cancer
- *ERCC2* mutant bladder cancer has altered genomic distribution of somatic mutations
- CTCF-cohesin binding sites are mutation hotspots in *ERCC2* mutant bladder cancer
- Somatic mutation distribution distinguishes passenger and driver *ERCC2* mutations

Authors

Jayne A. Barbour, Tong Ou,
Haocheng Yang, ..., Nikola A. Bowden,
Song Wu, Jason W.H. Wong

Correspondence

wusong@szu.edu.cn (S.W.),
jwhwong@hku.hk (J.W.H.W.)

In brief

ERCC2 driver mutations in bladder cancer are associated with cisplatin sensitivity, but their effect on genome instability and prognosis has not been clarified. In their recent reanalysis of 382 whole-genome-sequenced bladder cancers, Barbour et al. find that *ERCC2* mutations cause substantial alterations to genome-wide patterns of somatic mutations.

Barbour et al., 2024, Cell Genomics *4*, 100627
August 14, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2024.100627

Cell Genomics

CellPress
OPEN ACCESS

Article

ERCC2 mutations alter the genomic distribution pattern of somatic mutations and are independently prognostic in bladder cancer

Jayne A. Barbour,[1] Tong Ou,[2] Haocheng Yang,[1] Hu Fang,[1,3] Noel C. Yue,[1] Xiaoqiang Zhu,[1] Michelle W. Wong-Brown,[4,5] Yuen T. Wong,[6] Nikola A. Bowden,[4,5] Song Wu,[2,7,*] and Jason W.H. Wong[1,8,9,10,*]

[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[2]Urology Institute of Shenzhen University, The Third Affiliated Hospital of Shenzhen University, Shenzhen University, Shenzhen, China
[3]Institute of Biomedical Data, South China Hospital, Medical School, Shenzhen University, Shenzhen, China
[4]Centre for Drug Repurposing and Medicines Research, University of Newcastle, Newcastle, NSW, Australia
[5]Hunter Medical Research Institute, Newcastle, NSW, Australia
[6]Adult Cancer Program, Lowy Cancer Research Centre, UNSW, Sydney, NSW, Australia
[7]Department of Urology, South China Hospital, Medical School, Shenzhen University, Shenzhen, China
[8]Centre for Oncology and Immunology, Hong Kong Science Park, Hong Kong SAR, China
[9]Centre for PanorOmic Sciences, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
[10]Lead contact
*Correspondence: wusong@szu.edu.cn (S.W.), jwhwong@hku.hk (J.W.H.W.)

https://doi.org/10.1016/j.xgen.2024.100627

SUMMARY

Excision repair cross-complementation group 2 (*ERCC2*) encodes the DNA helicase xeroderma pigmentosum group D, which functions in transcription and nucleotide excision repair. Point mutations in *ERCC2* are putative drivers in around 10% of bladder cancers (BLCAs) and a potential positive biomarker for cisplatin therapy response. Nevertheless, the prognostic significance directly attributed to *ERCC2* mutations and its pathogenic role in genome instability remain poorly understood. We first demonstrated that mutant *ERCC2* is an independent predictor of prognosis in BLCA. We then examined its impact on the somatic mutational landscape using a cohort of *ERCC2* wild-type (n = 343) and mutant (n = 39) BLCA whole genomes. The genome-wide distribution of somatic mutations is significantly altered in *ERCC2* mutants, including T[C>T]N enrichment, altered replication time correlations, and CTCF-cohesin binding site mutation hotspots. We leverage these alterations to develop a machine learning model for predicting pathogenic *ERCC2* mutations, which may be useful to inform treatment of patients with BLCA.

INTRODUCTION

Excision repair cross-complementation group 2 (*ERCC2*) encodes xeroderma pigmentosum group D (XPD), a 5′-3′ helicase that is a component of the transcription factor II H (TFIIH) protein complex. TFIIH plays essential roles in transcription initiation through its interaction with RNA polymerase II subunit A and nucleotide excision repair (NER) when recruited to damaged lesions. Compound heterozygous mutations in *ERCC2* can cause the recessive genetic disorders xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy, which typically present with ultraviolet (UV) light sensitivity due to deficiencies in NER function.[1] These compound heterozygous mutations include the complete loss of function in one allele and a less deleterious point mutation in the other. The location of the *ERCC2* point mutations can vary but have been shown to affect XPD's helicase activity, stability, and interactions with other TFIIH proteins.[2,3] Somatic missense mutations in XPD are also putative drivers in cancers, with ∼10% of bladder cancer (BLCA) samples harboring these alterations.[4,5] *ERCC2* BLCA mutations do not overlap those underlying genetic disorders but are commonly found in the helicase domains of XPD. *ERCC2* mutant BLCAs are sensitive to cisplatin therapy, indicating a reduced capacity for the repair of cisplatin adduct DNA lesions, implying a deficiency in NER of these samples.[4,6,7] Clinically, although *ERCC2* mutation status has also been demonstrated to be a marker for good prognosis in BLCA,[8] it has never been shown to be an independent predictor due to insufficient cohort size.

While experimental evidence points to *ERCC2* mutations leading to NER deficiency, its functional impact on cancer development remains unclear. *ERCC2* mutant BLCA has previously been associated with the enrichment of the mutational signature SBS5,[9] but how mutant XPD causes this mutational signature is unknown. The other major mutational process occurring in BLCA can be attributed to the nucleic acid editing enzyme apolipoprotein B mRNA editing catalytic polypeptide-like family (APOBEC).[10] APOBEC is a cytosine deaminase that deaminates cytosine to uracil, causing C>T mutations targeted at viral RNA, but it can also affect host DNA.[10] Whether *ERCC2* mutations interact with APOBEC-associated mutagenesis has not been examined. Furthermore, somatic mutation density can be highly varied across the genome, correlating strongly with various epigenetic features, including chromatin accessibility,[11] histone modifications,[12,13] transcription factor binding,[14,15] and cytosine methylation.[16] How mutant XPD affects the distribution of somatic mutations across the genome remains to be explored.

In this study, we sought to determine how hotspot mutations in *ERCC2* are associated with changes in the distribution of mutations across the genome, with the aim to improve our functional understanding of how *ERCC2* mutations affect global mutagenesis and identify genomic characteristics that will help differentiate driver and passenger *ERCC2* mutations. We first showed that *ERCC2* mutation status is indeed an independent predictor of prognosis. To gain an insight into the *ERCC2*-mutant-driven mutational process, we compared the mutation distribution of *ERCC2* wild-type (WT) and mutant BLCA across a range of genetic and epigenetic features. *ERCC2* somatic mutations alter the mutational landscape of a range of mutational processes, with evidence implicating XPD in the repair of genomic uracil. We applied a machine learning approach to use the distribution of somatic mutations to differentiate driver and passenger *ERCC2* mutations in patients with BLCA, potentially enabling genomics-driven patient stratification for prognosis and platinum therapy.

RESULTS

Mutant *ERCC2* is an independent predictor of favorable prognosis in BLCA

Previous studies have shown that *ERCC2* mutation status predicts platinum sensitivity and is associated with good prognosis in patients with BLCA. However, these studies were carried out in relatively small cohorts (<100 patients). As *ERCC2* mutation status correlates with earlier tumor stage and tumor mutation burden, both of which are also associated with good prognosis, further analysis is required to establish the independent clinical significance of *ERCC2* mutation status.

We reanalyzed mutation data from a previously published series of 1,244 patients with bladder urothelial cancer sequenced using the MSK-IMPACT assay, including mutational profiling of the *ERCC2* gene.[17] In total, 156 patients were found to have an *ERCC2* mutation, while 1,088 patients were WT. Of the 156 mutant samples, 134 were missense at recurrent hotspots (see STAR Methods), 11 were other missense mutations, and 11 were nonsense or splice site mutations. As the pathogenicity of non-recurrent missense, nonsense, and splice site mutations is uncertain, only *ERCC2* missense at recurrent hotspots (labeled as mutant) and WT samples were retained. Finally, 113 *ERCC2* mutant and 886 *ERCC2* WT samples with complete age, mutation count, and specimen stage information were used for analysis.

Consistent with previous findings, *ERCC2* mutants had a significantly better prognosis than WT (p = 0.0023, log-rank test, Figure 1A). Proportionately, *ERCC2* mutants had significantly more early-stage (I-II) BLCA samples than WT (30.97% vs. 24.38%, p < 0.0175, Fisher's exact test, Figure 1B). The tumor mutation burden of *ERCC2* mutants was significantly higher than WT samples (median 27 vs. 9, p < 0.0001, Mann-Whitney test, Figure 1C), also consistent with previous findings.[9]

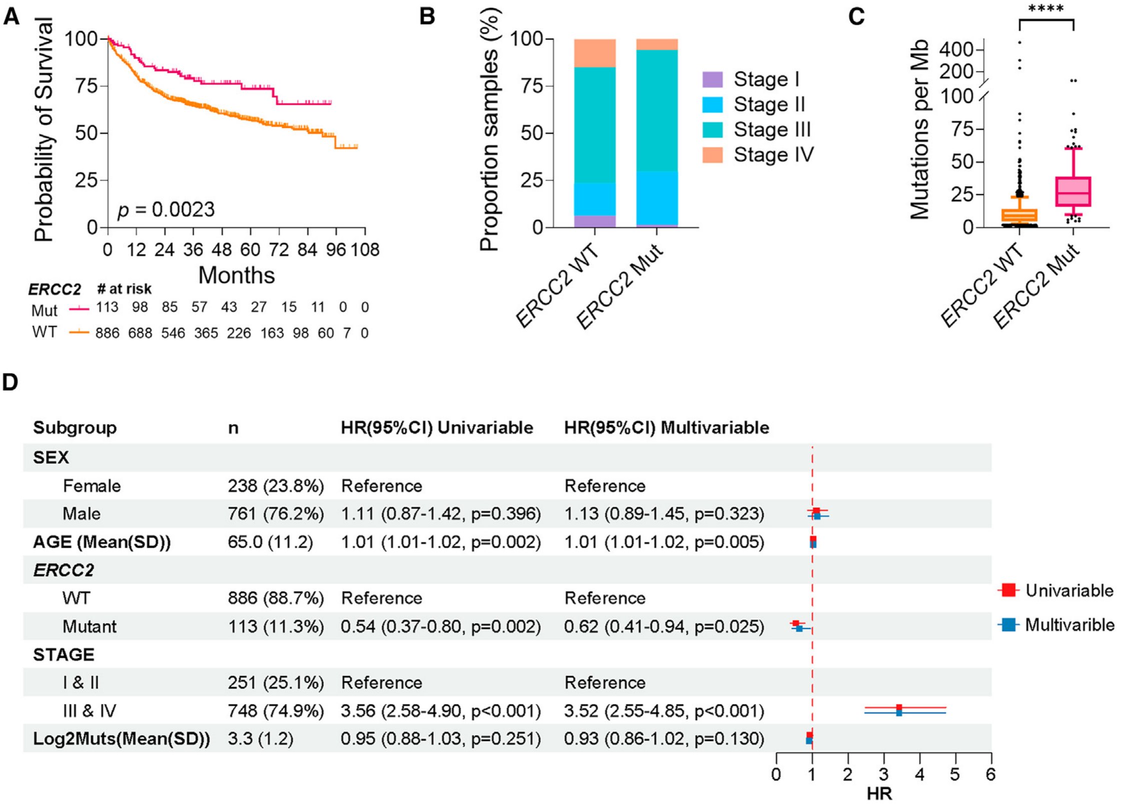
To resolve whether *ERCC2* mutation status is an independent predictor of prognosis in BLCA, we performed multivariable Cox regression, adjusting for sex, age, tumor stage, and tumor mutation burden. Despite strong correlations with tumor mutation burden and tumor stage, patients with mutant *ERCC2* independently have a significantly better prognosis than WT (hazard ratio [HR] = 0.62, p = 0.025, Figure 1D). To further demonstrate the independence of the *ERCC2* mutation and tumor mutation burden on prognosis, we selected patients with top and bottom quartile mutation counts and then stratified both these groups based on *ERCC2* mutation status, such that the difference in mutation count between the *ERCC2* mutant and WT is not significantly different within the top and bottom quartiles (Figure S1A). For *ERCC2* WT samples, there was no significant difference in prognosis between high- and low-mutation-burden tumors (p = 0.8824, log-rank test, Figure S1B). *ERCC2* mutants with high mutation burden had significantly better prognoses than both *ERCC2* WT groups (p = 0.0191 vs. WT low and p = 0.12 vs. WT high, log-rank test, Figure S1B). Only 3 *ERCC2* mutants had a low mutation burden, and the prognosis of these patients was worse than that of the other groups (Figure S1B).

As *ERCC2* is co-mutated with several genes at moderately high frequency (Figure S1C), we further sought to explore whether co-mutations may confound the independent influence of *ERCC2* mutations on prognosis. Of the 6 genes with a co-mutation frequency of >30% with *ERCC2*, *TP53* and *KDM6A* mutations are significantly associated with poorer (HR: 1.35 [1.10–1.66], p = 0.04, Cox's regression) and better prognoses (HR: 0.77 [0.61–0.96], p = 0.019, Cox's regression), respectively (Figure S1D). When combined with *ERCC2* mutations, it is evident that patients with *ERCC2* mutations always have a better prognosis than *ERCC2* WT, regardless of the effect of the co-mutated gene (Figures S1E–S1J). Thus, although certain rarer co-mutations may still influence the prognostic effect of *ERCC2* mutations, our results strongly support it as an independent factor in determining patient prognosis in BLCA.

Variable contribution of APOBEC-associated and other mutations in a cohort of 392 WGS BLCA samples

ERCC2 mutations have been linked to a specific mutational signature in BLCA,[9] but the genome-wide distribution of mutations associated with *ERCC2* mutants is unknown. To investigate this, we utilized the Genomics England (GE) cohort[18] of whole-genome-sequenced (WGS) BLCA and characterized samples that harbored putative *ERCC2* driver mutations. Out of 392 samples, 39 were characterized as *ERCC2* mutant and 343 as WT due to the complete absence of protein-altering *ERCC2* mutations. A further 10 samples were excluded from initial analyses, as they harbored a non-recurrent, protein-altering mutation in *ERCC2* that we could not confidently assign as either *ERCC2* mutant or WT (see STAR Methods).

To ensure that the phenotype responsible for these 39 samples is due to *ERCC2*, we compared the proportion of *ERCC2* mutant and WT samples with protein-altering mutations in cancer drivers. As with the MSK cohort, no other oncogene besides *ERCC2* was



Subgroup	n	HR(95%CI) Univariable	HR(95%CI) Multivariable
SEX			
Female	238 (23.8%)	Reference	Reference
Male	761 (76.2%)	1.11 (0.87-1.42, p=0.396)	1.13 (0.89-1.45, p=0.323)
AGE (Mean(SD))	65.0 (11.2)	1.01 (1.01-1.02, p=0.002)	1.01 (1.01-1.02, p=0.005)
ERCC2			
WT	886 (88.7%)	Reference	Reference
Mutant	113 (11.3%)	0.54 (0.37-0.80, p=0.002)	0.62 (0.41-0.94, p=0.025)
STAGE			
I & II	251 (25.1%)	Reference	Reference
III & IV	748 (74.9%)	3.56 (2.58-4.90, p<0.001)	3.52 (2.55-4.85, p<0.001)
Log2Muts(Mean(SD))	3.3 (1.2)	0.95 (0.88-1.03, p=0.251)	0.93 (0.86-1.02, p=0.130)

Figure 1. Association of *ERCC2* mutation status with clinical features in MSK BLCA cohort

(A) Kaplan-Meier estimate for *ERCC2* mutant and wild-type (WT) samples.

(B) Distribution of tumor stage in *ERCC2* mutant and WT samples.

(C) Boxplot of tumor mutation burden (mutation count) in *ERCC2* mutant and WT samples (****$p < 0.0001$, Mann-Whitney test).

(D) The hazard ratio (HR) and statistical significance using univariable and multivariable Cox regression models predicting survival based on sex, age, *ERCC2* mutation status, tumor stage, and mutation count (Log_2Muts).

See also Figure S1.

exclusively found in the *ERCC2* mutant samples (Figure S2A). Previous studies have found the presence of strong APOBEC mutational signatures in many BLCA samples[10] and increased signature 5 (SBS5), specifically in *ERCC2* mutant BLCA samples, but the analysis was restricted to exomes.[9] Since APOBEC-related mutations are frequent in BLCA and have a distinct mutational process, we first assessed the contribution of APOBEC and non-APOBEC-related processes (other) in *ERCC2* mutant and WT BLCA, where APOBEC-associated mutations are those in the T[C>R]N context (Figure 2A). Compared with WT samples, *ERCC2* mutant samples had higher and lower contributions to SBS2 and SBS13, respectively ($p < 0.0001$ and $p = 0.0005$, respectively, unpaired t test, Figure 2B), while, in line with previous findings, *ERCC2* mutants had a higher contribution of SBS5 compared with WT samples, but this did not reach significance in this cohort ($p = 0.063$, unpaired t test, Figure 2B). However, a significant difference in contribution to SBS1 was observed ($p < 0.0001$, unpaired t test, Figure 2B).

As APOBEC is a highly distinctive mutational process, for further analyses, we split APOBEC and other mutations (Figure 2C). For confirmation, APOBEC mutations were more similar to SBS2 and SBS13, while other mutations were more similar to SBS5 (Figure 2D). We next investigated the distribution of APOBEC and other mutations across the genome by calculating the observed/expected mutation ratio in 1 Mb windows. For an illustration, the variation in the mutation distribution of chromosome 1 is shown in Figure 2E. Based on the principal-component analysis of the observed/expected mutation ratio of these windows, the genome-wide distribution of other and APOBEC mutations in WT and *ERCC2* mutant samples could be readily distinguished (Figure 2F).

ERCC2 mutant samples display altered genomic distribution of somatic mutations

Mutations in most mismatch-repair-proficient cancers show variation in mutation burden in relation to replication timing, chromatin accessibility, and gene expression.[11] To further explore the relationship between mutation density and these epigenomic features, mutation densities for APOBEC and other mutations were calculated for gene bodies and replication time in *ERCC2* mutant and WT BLCA. Expressing mutation density as the observed-expected mutation ratio (see STAR Methods), we found that the distribution of APOBEC and other mutations was significantly higher in all genic regions and lower in intergenic regions in *ERCC2* mutant samples compared with WT

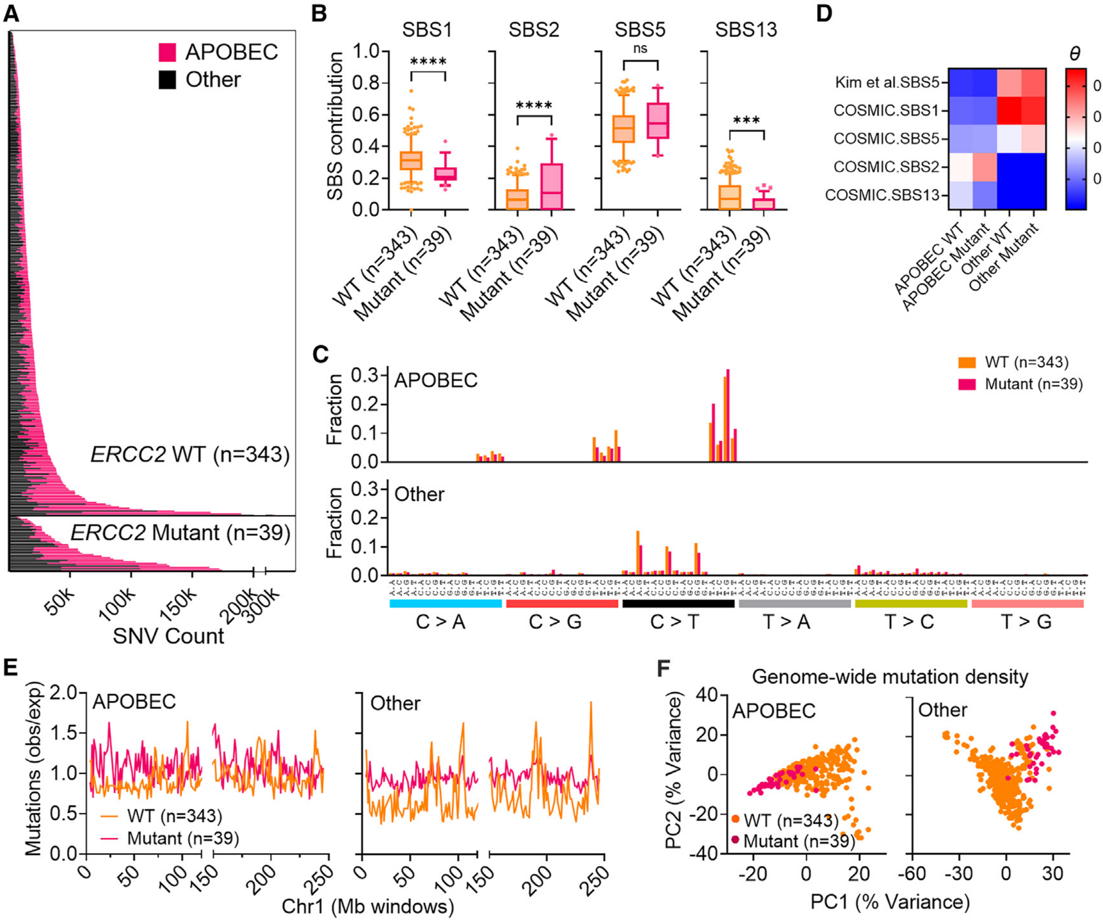


Figure 2. Contribution and distribution of APOBEC and other mutations in *ERCC2* mutant and WT BLCA

(A) Total number of mutations attributed to T[C>D]N (APOBEC) or not T[C>D]N (other) in GE WGS BLCA cohort arranged by genotype and total mutation number.

(B) Signature contribution (as a fraction calculated with deconstructSigs) of all SNVs in WT and *ERCC2* mutant GE BLCA samples for COSMIC signatures 1, 2, 5, and 13. Box and whiskers represent quartiles and 5th–95th percentile, respectively.

(C) Trinucleotide mutational spectra of APOBEC and other mutations for WT (orange) and *ERCC2* mutant (pink) GE BLCA samples.

(D) Heatmap of cosine similarities to mutational signatures in GE BLCA SNVs that were separated into APOBEC and other from (C), where theta = 1 is most similar. Signatures 1, 2, 5, and 13 are from COSMIC, and the other signature, TCGA.130.DFCI.MSK.50.signature5, is from the supplementary material from Kim et al.[9]

(E) Observed-expected mutation density ratios for 1 Mb windows of hg38 chromosome 1 for APOBEC and other SNVs.

(F) Principal-component analysis (PCA) plots representing PC1 and PC2 of observed-expected mutation density ratios were calculated across each 1 Mb window of hg38 genome wide for APOBEC SNVs and other SNVs. n.s., not significant, ***p = 0.005 and ****p < 0.0001, unpaired t test.

See also Figure S2.

(Figure 3A). We found an increase in the burden of APOBEC-related mutations in the 5′ UTR relative to what is expected by chance (observed-expected ratio >1) in both *ERCC2* mutant and WT groups (Figure 3A). The increased burden of APOBEC-related mutations in the 5′ UTR in both *ERCC2* mutant and WT groups (Figure 3A) is consistent with previous findings that APOBEC causes mutation clusters around the start of active genes, which could be from single-stranded DNA (ssDNA) exposure as transcription commences.[19] A linear regression between other mutation densities and replication time showed that WT samples had a slope of −0.01382 compared with −0.002548 for mutant (Figure 3B), with significantly decreased and increased burdens of mutations in mutant samples compared with WT in late- and early-replicating regions, respectively (q = 0.000086 and q < 0.000001, Student's t test with multiple testing correction, Figure S3A). APOBEC mutations had a striking trend of being negatively associated with replication time in WT samples (slope = −0.007405) but positively associated with replication time in *ERCC2* mutant samples (slope = 0.006413) (Figures 3B and S3A). The observed relationship between replication time and other mutations from *ERCC2* mutant BLCA is consistent across all mutation types (Figure S3B). The differential burden of other mutations between mutant and WT samples in gene bodies and over the replication time landscape suggested that transcriptionally active or open chromatin plays a role in the distribution of *ERCC2*-related mutagenesis.

We next examined the effect that transcriptional activity has on mutagenesis in the BLCA genomes. *ERCC2* mutant samples

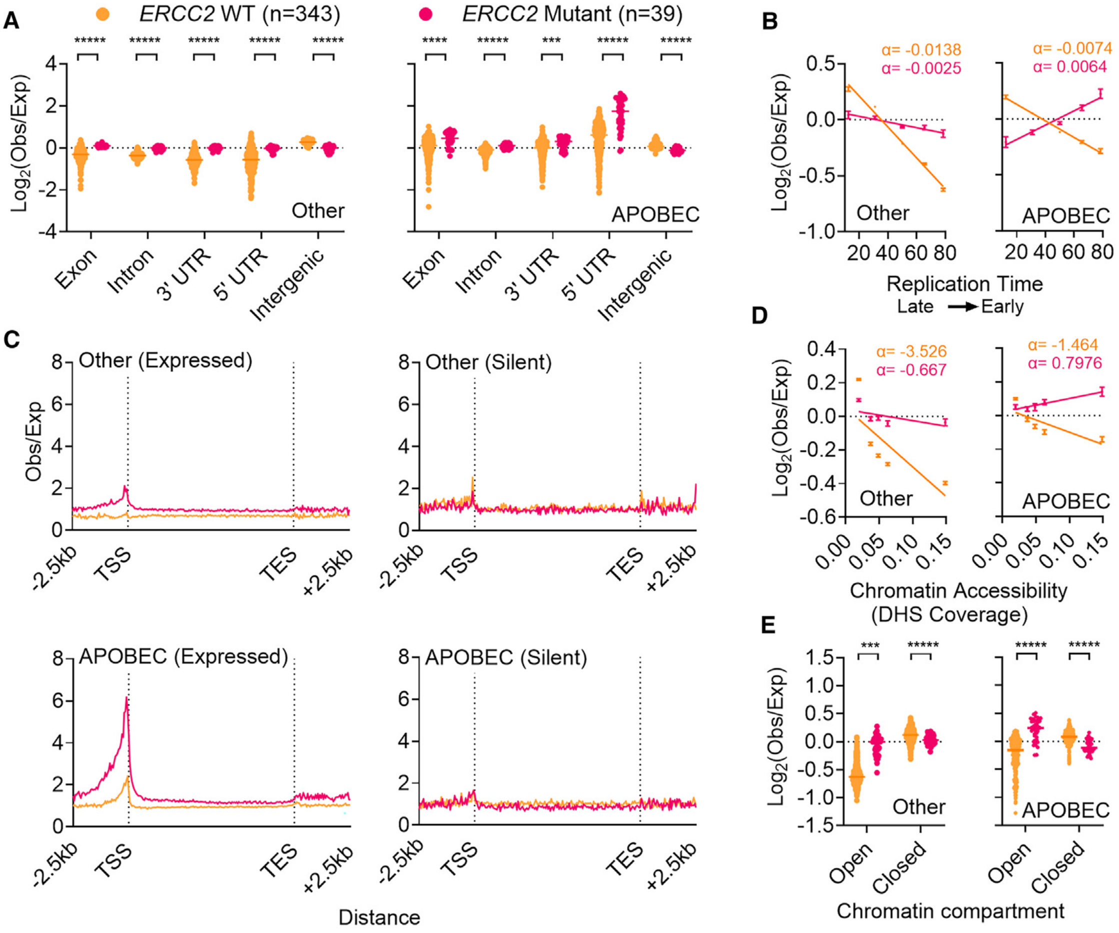


Figure 3. Genome-wide distribution of APOBEC and other mutations in *ERCC2* mutant and WT BLCA

(A) Mutation densities as $\log_2$ observed-expected ratios ($\log_2$(obs/exp)) in exons, and 3′ and 5′ UTRs, introns, or not in any of these regions (intergenic) in WT samples in orange and *ERCC2* mutant samples (mutant) in pink, with other SNVs displayed on the left and APOBEC SNVs displayed on the right. Median is indicated by a solid line.

(B) Mutation densities as $\log_2$(obs/exp) for 5 genomic bins organized by replication time. BrdU immunoprecipitation coverage was used for binning, where a higher number indicates an earlier replication time. Plots and error bars represent mean and standard deviation of different samples, and the line represents a linear regression model between mutation densities and the mean replication time for each of the bins. Points represent the mean with standard error.

(C) Observed-expected mutation density ratio profile plots of other SNVs (left) and APOBEC SNVs (right) across gene bodies of genes expressed in bladder tissue (expressed genes) or genes not expressed in bladder tissue (silent genes). TSS, transcriptional start site; TES, transcriptional end site. The gene body was organized into 150 bins, and the region 2.5 kb up- or downstream of the TSS or TES was organized into 50 bins.

(D) Mutation densities as $\log_2$(obs/exp) for 5 genomic bins organized by chromatin accessibility measured by DNase hypersensitive coverage, where a higher number is more accessible. Plots and error bars represent mean and standard deviation of different samples, and the line represents a linear regression model between mutation densities and the mean DHS coverage for each of the bins. Points represent the mean with standard error

(E) Mutation densities as $\log_2$(obs/exp) mutation density ratios for genomic regions previously annotated in normal bladder to be either a chromatin A compartment (open) or chromatin B compartment (closed).[42] ***q < 0.0001 and *****q < 0.000001, Student's t test with multiple testing correction. Median is indicated by a solid line.

See also Figure S3.

have an increased genic mutation burden for other and APOBEC mutations compared to WT, specifically at expressed genes (Figure 3C). This was particularly pronounced immediately before the transcriptional start site (Figures 3C and S3C).

To look more generally at active and inactive chromatin genome wide, we next used DNase I hypersensitivity (DHS) to compare the burdens of APOBEC and other mutations between *ERCC2* mutant and WT samples. Regression was performed between DHS coverage and somatic mutation densities across 5 bins, and we found a marked difference in the relationship between *ERCC2* mutant and WT cancer (Figure 3D). *ERCC2* mutant and WT samples had slopes of −3.526 and −0.6667, respectively, for other SNVs and −1.464 and 0.7976 for APOBEC SNVs. For APOBEC SNVs, there were significantly more mutations in *ERCC2* mutants compared with WT in most accessible DHS regions and significantly less in less accessible regions (q < 0.000001, Student's t test with multiple testing correction, Figure 3D). We also found significantly increased

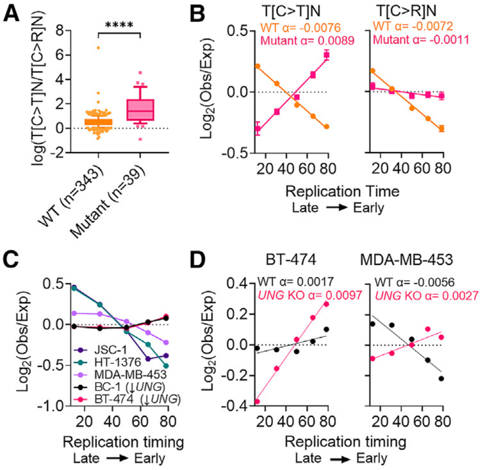


Figure 4. APOBEC mutagenesis in WT and *ERCC2* mutant cancer

(A) The log ratio of APOBEC mutations that can be attributed to T[C>T]N or T[C>R]N, where R is A or G. Box and whiskers represent quartiles and 10th–90th percentile, respectively.

(B) Mutation densities for T[C>T]N (SBS2) and T[C>R]N (SBS13) mutations as $\log_2$(obs/exp) for 5 genomic bins organized by replication time for *ERCC2* mutant and WT BLCA. Points represent the mean with standard error.

(C) Mutation densities as $\log_2$(obs/exp) for 5 genomic bins organized by replication time for the following cell lines that contain endogenous APOBEC mutational signatures: BC-1, BT-474, HT-1376, JSC-1, and MDA-MB-453. Of these cell lines, BC-1 and BT-474 have low UNG expression.

(D) Mutation densities as $\log_2$(obs/exp) for 5 genomic bins organized by replication time for WT and *UNG* CRISPR knockout in BT-474 and MDA-MB-453 cells. Mutation data for cell lines are from Petljak et al.[21] ****$p < 0.00001$ by unpaired t tests.

See also Figure S3.

and decreased burdens of other mutations in *ERCC2* mutant samples in open compartments and closed compartments, respectively, compared with WT ($q < 0.000001$ and $q = 0.000024$, Student's t test with multiple testing correction, Figure 3E). The exclusion of CTCF-cohesin binding sites (CBSs) and genes did not affect the observed associations (Figures S3D and S3E).

Earlier, we had shown that SBS1 was a significant contributor to many samples (Figures 2B and 2C). We further separated other mutations into [C>T]pG mutations and other remaining (i.e., non-[C>T]pG) mutations. However, we did not find this to increase the distinction between WT and *ERCC2* mutant samples in relation to replication timing and DNase hypersensitivity (Figures S3F and S3G). Together, our results provide evidence that *ERCC2* protects accessible chromatin from mutagenesis.

ERCC2 mutant cancers are enriched in APOBEC-induced cytosine-deamination-associated C>T mutations

APOBEC mutational signatures can be further divided into SBS2 and SBS13, where SBS2 is dominated by T[C>T]N and SBS13 is dominated by T[C>R]N (where R is A or G). Earlier, we showed that *ERCC2* mutant BLCAs have higher signature contributions to SBS2 compared with WT BLCAs, and the converse for SBS13 (Figure 2B). The differences in the distribution of APOBEC mutations across the replication time landscape were also particularly striking, with WT having a slope of −0.007405, while *ERCC2* mutants had a slope of 0.006413 (Figure 3B). We next explored APOBEC-related mutagenesis in more detail by separating mutations associated with SBS2 and SBS13 for further analysis.

The proportion of APOBEC mutations being T[C>T]N (i.e., SBS2) relative to T[C>R]N (i.e., SBS13) was significantly higher in *ERCC2* mutants compared with WT ($p < 0.001$, Student's t test, Figure 4A). APOBEC-associated T[C>T]N is known to be caused by mutations resulting from unrepaired U>G mismatches resulting from APOBEC-induced cytosine deamination to uracil.[10] APOBEC overexpressing yeast that are WT for uracil-DNA glycosylase (*UNG*) exclusively acquire C>D somatic mutations, whereas APOBEC overexpressing yeast that are also *UNG* mutants exclusively acquire C>T somatic mutations.[20] A recent study also found that human cancer cell lines with strong endogenous APOBEC signatures were enriched in SBS2 when they were *UNG* deficient.[21]

We next examined the relationship between APOBEC-associated mutations and replication time. We found that T[C>T]N mutations had a positive relationship with replication time in *ERCC2* mutant samples (slope = 0.008926) but a negative relationship with replication time in WT samples (slope = −0.007569) (Figure 4B). T[C>R]N mutations, on the other hand, display a less prominent difference in their relationship with replication time between WT and *ERCC2* mutants, with slopes of −0.007186 and −0.001147, respectively (Figure 4B).

We reanalyzed mutation patterns in previously published cancer cell line sequencing data with APOBEC signatures[21] for comparison to WT and *ERCC2* mutant BLCA to dissect APOBEC mutagenesis processes. BC-1 and BT-474 cells, which have low *UNG* expression, both displayed a positive relationship between T[C>T]N mutations and replication time, while the other *UNG*-proficient cell lines showed a strong negative trend (Figure 4C), corresponding to the *ERCC2* mutant and WT profiles, respectively. The study also profiled somatic mutations accumulated by cells with knockout (KO) of genes related to the APOBEC pathway, including *UNG*. Both BT-474 and MDA-MB-453 KO cell lines displayed a redistribution of T[C>T]N mutations toward a positive association with replication time (Figure 4D). Thus, APOBEC-associated mutations in *ERCC2* mutant BLCA share characteristics of mutations associated with cytosine-deamination-induced genomic uracil in *UNG*-deficient cells, suggesting that *ERCC2* may have a role in genomic uracil repair.

ERCC2 mutant cancers display strong mutation hotspots in CBSs

DHSs are associated with *cis*-regulatory elements, including promoters, enhancers, and CBSs. As mutations in *ERCC2* mutants showed increased mutation density at DHS regions (Figure 3D), we examined the different classes of *cis*-regulatory regions. We found striking mutation hotspots in *ERCC2* mutants at CBSs, with only a minor increase in mutation rate at promoters or enhancers (Figure 5A). We similarly observed CBS hotspots in a total of 7 *ERCC2* mutant BLCA samples from three other studies[22–24] (Figure S4A), as well as 3 *ERCC2* mutant liver cancer samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium[22] (Figure S4B).

While somatic mutation hotspots in CBSs have previously been reported in UV-associated skin cancers[25] and SBS17-associated gastrointestinal cancers,[26–28] it is a striking and novel observation that *ERCC2* mutant BLCAs also have CBS mutation hotspots. We also observed increased mutation densities in flanking regions in *ERCC2* mutant samples compared with WT, which may generally reflect greater chromatin accessibility of the CBS flank but is substantially lower compared with the CBS itself (Figure 5B). APOBEC SNVs also displayed elevated mutation densities at CBSs for *ERCC2* mutant, but not WT, BLCA (Figure S4C). However, these elevated mutation density profiles across the CTCF motif were not as striking in terms of observed/expected ratios as other SNVs (Figure 5A). To determine if CTCF binding directly affects mutagenesis at CBSs, we separated CBSs into quartiles by CTCF chromatin immunoprecipitation sequencing (ChIP-seq) signal. We found that the mutation density increased with an increasing CTCF signal (Figure 5C, $p < 0.001$, paired t test comparing mutation density across quartiles in each patient), and conversely, mutated sites also had a higher CTCF signal compared with non-mutated sites (Figure 5D, $p < 0.0001$, Mann-Whitney test), suggesting that CTCF binding is indeed important. Previous reports of CBS hotspots found specific mutational patterns and signatures across the CBS motif.[25,26] We observed that the CBS mutations in *ERCC2* mutants also have a similar distribution across the CTCF motif compared with CBS hotspots found in esophageal adenocarcinoma (ESAD) (Figures 5E and 5F), with a correlation coefficient of 0.925, but is different from melanoma (Figure 5G), which had a correlation coefficient of −0.034. This potentially implicates a shared mechanism of CBS mutagenesis with ESAD. In terms of the type of the CBS-specific trinucleotide mutational spectrum, there is strong enrichment for T>N mutations, with the strongest enrichment being T>G, which is absent from the CBS flank (Figure 5H). This is similar to gastroesophageal cancers with SBS17, where predominantly T>G and T>C mutations accumulate at CBSs.[26,27] For CBS enrichment, we further separated other mutations to [C>T]pG and remaining non-[C>T]pG mutations. Those that are non-[C>T]pG had a log2 mean of the difference between WT and *ERCC2* mutant groups of 3.27, but this was just 0.87 for C[C>T]pG mutations (Figure S4D), providing more evidence that T>G and T>C are the key mutational processes associated with CBS hotspots.

We next explored whether XPD may engage at CBSs. Using previously published XPD ChIP-seq data,[29] we found a strong enrichment of XPD at CBSs, with XPB also enriched (Figure 5I), implying that these proteins are co-bound to CBSs as part of the TFIIH complex. DNA binding was observed at both genic and intergenic CBSs (Figure 5I), suggesting that binding is independent of transcriptional activity. To investigate this further, we used TFIIH excision repair-sequencing (XR-seq) data from Hu et al.,[30] which measures TFIIH repair of cisplatin damage. *ERCC2* mutant BLCA samples lost the negative relationship between TFIIH repair and mutation burden, especially for other mutations (Figures S4E and S4F). This supports the hypothesis that defective repair is responsible for the mutation redistribution in *ERCC2* BLCA. Together, these results illustrate that the presence of XPD in the TFIIH complex may play a role in the maintenance of CBSs.

Genomic uracil accumulates at CBSs

Earlier, we found that *ERCC2* mutants may be associated with the dysfunctional repair of genomic uracil that results from APOBEC-associated cytosine deamination in BLCA. To determine if genomic uracil might also be associated with CBS mutagenesis, we took advantage of previously published genome-wide maps of genomic uracil in *UNG* KO cells (UdgX cross-linking and polymerase stalling sequencing [Ucaps-seq]).[31] Since Ucaps-seq is at base-pair resolution, uracil from incorporation and deamination can be identified based on whether the reference base is A/T or C/G, respectively. We found a strong enrichment of the incorporation of genomic uracil at CBSs (Figure 6A). As with the mutations, the distribution of uracil across the motif is asymmetric, mirroring *ERCC2* mutant CBS mutation hotspots (Figure 6B) with a correlation coefficient of 0.835. Treatment with pemetrexed, which increases uracil misincorporation, results in an even greater enrichment of uracil at CBSs (Figure 6A). Analysis of the trinucleotide context of uracil misincorporation shows that the frequency of uracil incorporation sites was most enriched in TTT, followed by CTT (Figure 6C). This resembles trinucleotides most strongly mutated in the CBS motif in *ERCC2* mutant cancers and is also the most prevalent trinucleotide context in SBS17, where CBS hotspots are also observed.[16,17,20] In contrast, uracil from cytosine deamination did not share the same trinucleotides as APOBEC mutations (Figure S5), but this may be because the Ucaps-seq data were acquired in HeLa cells, which do not constitutively express active APOBEC.

To further confirm the relationship between CTCF binding and the presence of genomic uracil, we stratified CBSs into quartiles based on ChIP-seq/control signal and calculated the normalized uracil count in each quartile. We observed an increasing trend of higher uracil content for CBSs with higher CTCF binding (Figure 6D). Consistently, CBSs with enrichment of uracil (i.e., dU-input count >0) had significantly stronger CTCF binding compared with those without (Figure 6E, $p < 0.0001$, Mann-Whitney test). Together, these findings suggest that the CBS mutational hotspots in *ERCC2* mutants may result from misincorporation of uracil and subsequent error-prone DNA repair.

Genome-wide distribution of somatic mutations predicts pathogenic *ERCC2* mutations in BLCA

Patients with cancer with tumors harboring *ERCC2* mutations have favorable responses to cisplatin.[4,6] *ERCC2* is included on most somatic targeted sequencing panels. However, the pathogenicity of *ERCC2* mutations is not always apparent, as they often occur outside of known hotspots.[6] We, therefore, wanted to test if it is possible to predict a sample's *ERCC2* mutation status based on its genome-wide distribution of somatic mutations with the idea that WGS data could be used to verify the pathogenicity of uncertain *ERCC2* mutations. To this end, we generated support vector machine (SVM) models to classify whether a sample is *ERCC2* mutant based on the local and global distributions of somatic mutations.

First, using leave-one-out cross-validation on the GE BLCA cohort with well-defined *ERCC2* mutation status (WT = 343, mutant = 39), the SVM models achieved 98.69% accuracy (sensitivity: 97.22%, specificity: 98.84%, Figure 7A). One of the

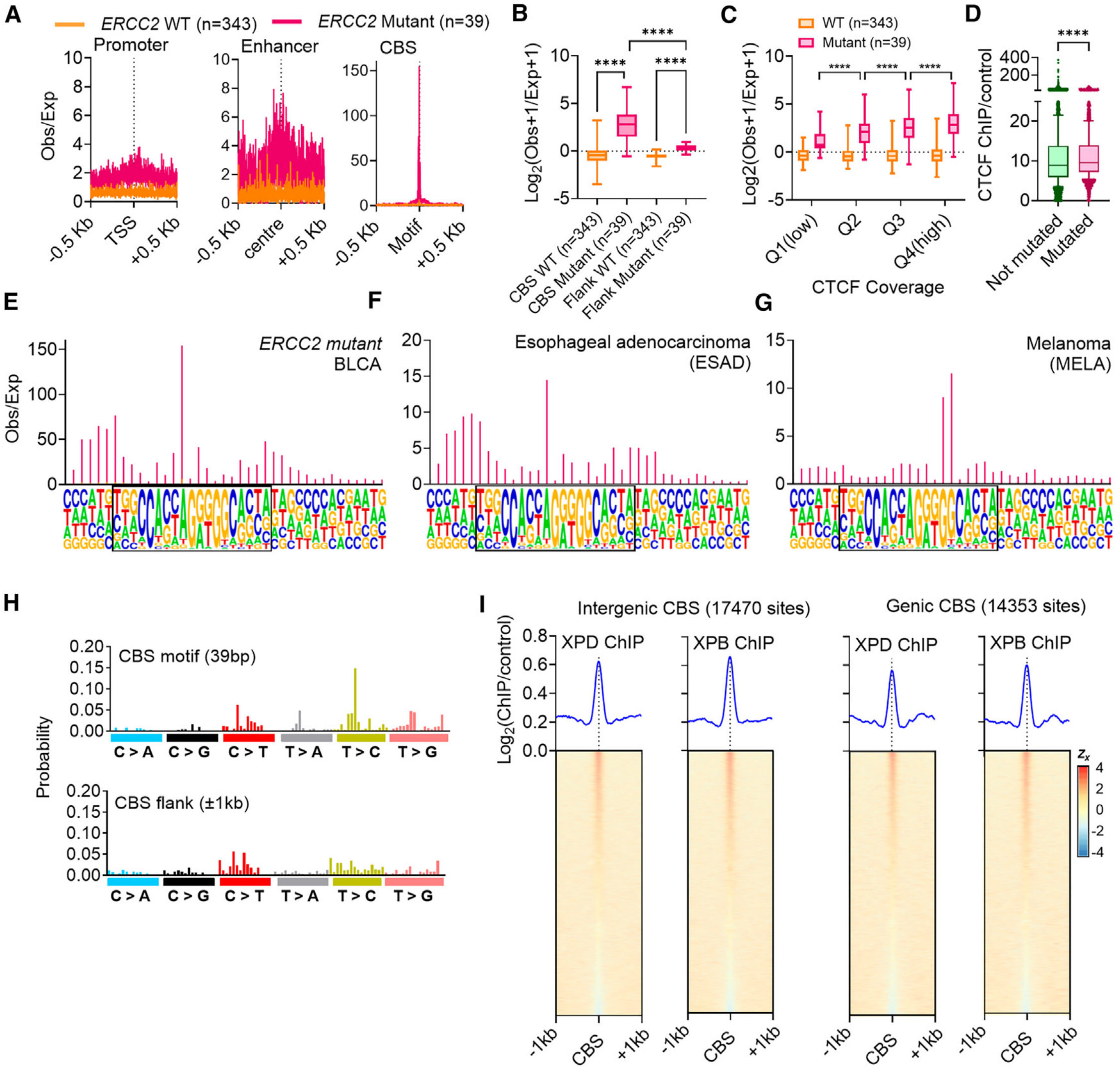


Figure 5. Mutation densities at DNase hypersensitive regions in *ERCC2* mutant and WT BLCA

(A) Profile plots of mutation densities as obs/exp for Other mutations in regions annotated in normal bladder as the promoter, enhancer, or CTCF-cohesin binding site (CBS) ± 0.5 kb of the TSS, center, and motif, respectively.

(B) Mutation densities as $\log_2$ (obs+1/exp+1) of CBS motif and ± 1 kb flanking regions for WT and *ERCC2* mutant GE BLCA samples. ****q < 0.0001, Student's t test with multiple testing correction. Box and whiskers represent quartiles and minimum-maximum, respectively.

(C) Mutation densities as $\log_2$ (obs+1/exp+1) at CBSs were organized into quartiles based on ENCODE kidney CTCF ChIP-seq coverage where Q1(low) and Q4(high) represent CBSs with the lowest and highest CTCF coverage, respectively. ****q < 0.0001, Student's t test with multiple testing correction. Box and whiskers represent quartiles and minimum-maximum, respectively.

(D) ENCODE kidney CTCF ChIP-seq coverage at CBSs with the presence (mutated) or absence of (not mutated) of other mutations from *ERCC2* mutant GE BLCA samples. ****$p < 0.0001$, Mann-Whitney test. Box and whiskers represent quartiles and 10th–90th percentile, respectively.

(E–G) Observed/expected mutational profile of GE BLCA *ERCC2* mutant (E), esophageal adenocarcinoma (F), and melanoma (G) samples across the CBS motif.

(H) Trinucleotide mutation frequencies of BLCA *ERCC2* mutations in CBS and flanking regions. This was normalized using manual normalization in deconstructSigs to account for the trinucleotide composition of the regions.

(I) Profile and heatmaps of coverage of XPD and XPB ChIP-seq across intergenic (left) and genic (right) CBS. Data were accessed from GEO: GSE44849.

See also Figure S4.

WT samples that was misclassified was predicted to have the highest probability of being a mutant. We manually examined the unfiltered VCF file for mutations in *ERCC2* in this sample and found that the sample, in fact, contains a N238T mutation at a relatively low variant allele frequency (4/84 reads), which was filtered by the variant caller. N238 is a known mutation hotspot in BLCA; thus, our classifier is also able to identify functional *ERCC2* mutations that are missed by variant callers. For comparison, we also trained an SVM model using only trinucleotide context mutations. The trinucleotide model was less accurate,

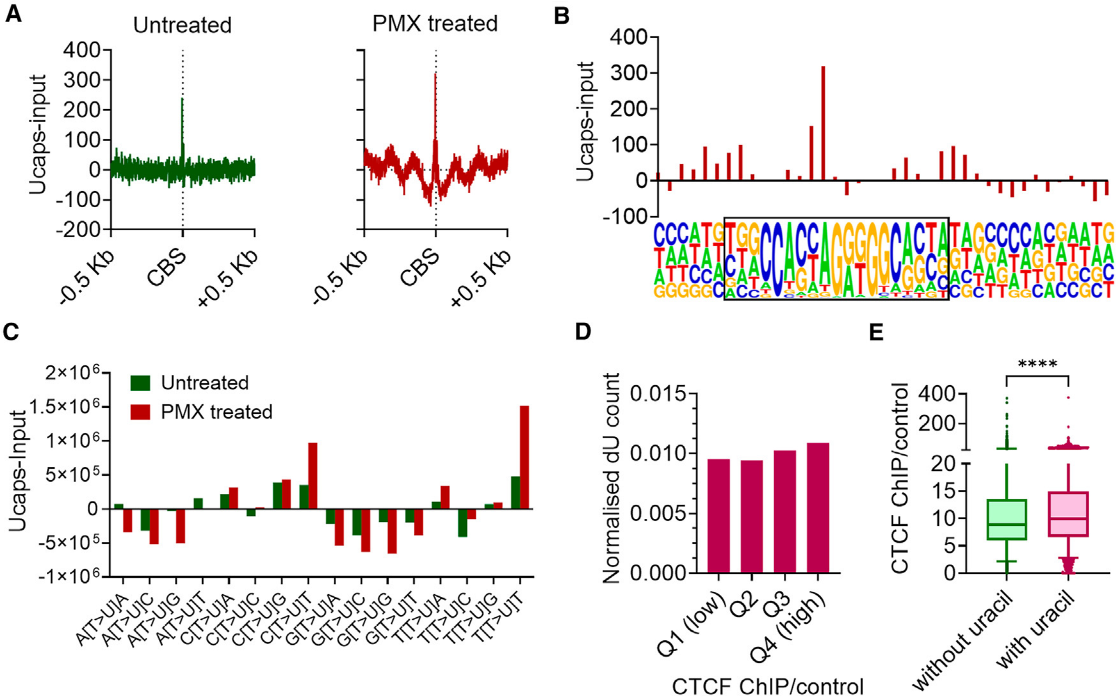


Figure 6. Genomic uracil distribution at CBSs

(A) Profiles of Ucaps-seq data (SRA: SRP319102) as input (inp) subtracted from experiment (exp) across CBSs with a 0.5-kb flank up- (+) and downstream (−).

(B) Ucaps-seq exp-inp across each base of the CTCF motif of CBSs.

(C) Frequency of single base sites of uracil incorporation in the trinucleotide context as Ucaps-seq exp-inp. T>U indicates that a uracil was detected where the reference base is a thymine.

(D) Uracil (dU) incorporation at CBSs organized into quartiles based on ENCODE kidney CTCF ChIP-seq coverage where Q1(low) and Q4(high) represent CBSs with the lowest and highest CTCF coverage, respectively. dU incorporation is normalized by the count of the number of dT and dA in each region based on the reference genome

(E) ENCODE kidney CTCF ChIP-seq coverage at CBSs with the presence of (with uracil) or absence of (without uracil) dU. Here, uracil presence was defined as a CBS with Ucaps-seq exp-inp ≥ 1 and absence as Ucaps-seq exp-inp ≤ 0. ****$p < 0.0001$, Mann-Whitney test. Box and whiskers represent quartiles and 5th–95th percentile, respectively.

See also Figure S5.

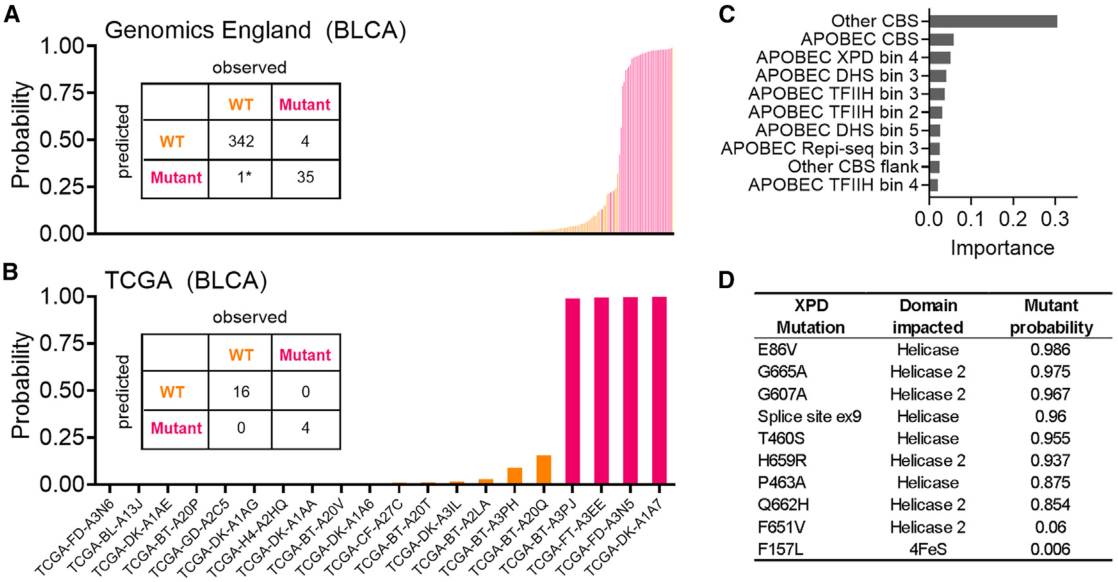
at 95.83% (sensitivity: 84.85%, specificity: 96.87%, Figure S6), but may nevertheless be valuable when whole-genome sequencing data are unavailable.

We further evaluated the SVM model trained on the GE cohort using an independent TCGA WGS BLCA cohort (WT = 19, mutant = 4), achieving 100% accuracy in classifying WT and mutant samples (Figure 7B). A sensitivity analysis was performed on the SVM model to determine the relative importance of the mutational features in predicting *ERCC2* mutation status. This found that the observed/expected ratio of Other mutations at CBS was by far most important, followed by the observed/expected ratio of APOBEC mutations at CBSs (Figure 7C), supporting our observation that CBS hotspots are highly distinctive in *ERCC2* mutant cancers.

Finally, we apply our SVM model to the GE BLCA cohort, which we had excluded from our earlier analysis, as their *ERCC2* mutation status is uncertain due to the mutations not being located at known hotspots (*n* = 10). The SVM predicted all but two samples to be *ERCC2* mutant. The two samples predicted to be WT included one with an F157L mutation and one with an F651V mutation (Figure 7D). F157L is outside the helicase domains, while F651V is not within a conserved helicase motif, which may explain the lack of functional effect in these samples.

DISCUSSION

In this study, we show that *ERCC2* mutation status is an independent prognostic factor in BLCA. We further show that these cancers display a distinctive genomic distribution of somatic mutations, including mutation hotspots at CBSs. We leverage this knowledge to build an SVM model that differentiates driver and passenger *ERCC2* mutations. This is particularly useful for mutations in *ERCC2,* as pathogenic mutations appear as point mutations that are widely distributed across the protein. Pathogenicity is currently inferred based on hotspot sites found across patients with BLCA. However, this means that pathogenic mutations appearing at rarer sites can be missed. Our SVM model not only classified the *ERCC2* mutation status of samples with 99% accuracy but also identified one sample where the *ERCC2* mutation was filtered due to low variant allele frequency. Using our SVM model, we further identified 8 samples with non-recurrent



		WT	Mutant
predicted	WT	342	4
	Mutant	1*	35

		WT	Mutant
predicted	WT	16	0
	Mutant	0	4

XPD Mutation	Domain impacted	Mutant probability
E86V	Helicase	0.986
G665A	Helicase 2	0.975
G607A	Helicase 2	0.967
Splice site ex9	Helicase	0.96
T460S	Helicase	0.955
H659R	Helicase 2	0.937
P463A	Helicase	0.875
Q662H	Helicase 2	0.854
F651V	Helicase 2	0.06
F157L	4FeS	0.006

Figure 7. SVM model predicts *ERCC2* mutation status based on genome-wide distribution of somatic mutations

(A) Classification of samples as WT (orange) or *ERCC2* mutant (pink) from SVM and associated confusion matrix showing the accuracy of predictions by leave-one-out cross-validation of the GE cohort.

(B) Classification of samples as WT (orange) or *ERCC2* mutant (pink) from SVM and associated confusion matrix showing the accuracy of predictions of PCAWG BLCA samples.

(C) Sensitivity analysis showing the importance of features as predictors in SVM model.

(D) Prediction from SVM when applied to 10 GE samples with non-recurrent but protein-altering mutations in *ERCC2*. Probability is the prediction from the model, and the domain impacted refers to the protein domain of XPD where the mutation is located.

See also Figure S6.

(from over 120,000 cancer samples, including over 3,000 BLCA samples) functional *ERCC2* mutations. Besides prognostic stratification, this WGS approach to detect pathogenic *ERCC2* mutations can have clinical significance, as patients with BLCA with these mutations are more sensitive to cisplatin treatment.[5,7] Our approach can be analogous to the use of somatic mutations to infer the homologous recombination deficiency score to predict response to PARP inhibitors[32] and other DNA-damaging agents,[33] an approach that is now commonly used in clinical settings.[33]

While the biological mechanisms underlying *ERCC2*-mutation-driven mutagenesis require further investigation, our results support the role of *ERCC2* mutants in compromised DNA repair at open chromatin. NER is more active at open chromatin,[34] and NER proteins *ERCC1*[35] and *ERCC6*[36] are located at and actively participate in DNA repair at CBSs, making it feasible that XPD (*ERCC2*) does as well. It is plausible that in *ERCC2* mutant BLCA, NER activity is lost, increasing the relative mutation burden in those regions. We also found that mutations from *ERCC2* mutant BLCA had an inverse relationship between mutation burden and TFIIH-mediated repair (Figures S4E and S4F), which supports a loss of NER activity in *ERCC2* mutant samples.

Analysis of mutations from cell lines with endogenous APOBEC signatures and KO of APOBEC-related genes revealed that the mutations of *ERCC2* mutant BLCA mirror those of *UNG*-deficient cell lines (Figures 4C and 4D) implying that the damage that *ERCC2* is repairing is uracil. *UNG* encodes a uracil excision enzyme (UDG2), implicating a role for *ERCC2* in the repair of genomic uracil. Interestingly, SBS17, which is also associated with mutation hotspots at CBSs, may also be caused by genomic uracil. SBS17 can be recapitulated in cell culture experiments by treatment with 5-FU, and SBS17 mutations develop in breast and colorectal cancer tumors when patients are treated with 5-FU.[37] dUTP can be incorporated into DNA by mammalian polymerases at similar efficiencies to dTTP, but the cell gets around this by keeping dUTP levels low. However, 5-FU is a thymidylate synthase inhibitor, which increases the dUTP/dTTP ratio in the cell, leading to genomic uracil misincorporation. Incorporation of dUTP can cause T>G mutations,[38–40] which are the hallmark of SBS17. UDG2 does not distinguish between dU generated by cytosine deamination attributed to APOBEC activity or erroneous misincorporation of dUTP from the free nucleotide pool, such as those from 5-FU treatment.[41] Therefore, these seemingly unrelated mutational processes of APOBEC cytidine deamination in BLCA and uracil misincorporation in SBS17 may be related by uracil excision, whether by deamination or misincorporation. As further evidence for the relationship between genomic uracil and SBS17, we also found that somatic mutation hotspots frequently found in cancers with SBS17 are also sites with increased accumulation of genomic uracil. We therefore hypothesize that *ERCC2* and UDG2 collaborate in the repair of genomic uracil. Given that uracil misincorporation is not necessarily mutagenic, mutant *ERCC2* may cause increased erroneous repair of abasic sites following the excision of uracil by UDG2. Another intriguing possibility is that *ERCC2* unwinds DNA around genomic uracil, as UDG2 has an order of magnitude higher activity against uracil in ssDNA than double-stranded DNA.[41] Further experiments will be required to fully elucidate the role of *ERCC2* in the repair of genomic uracil.

Limitations of study

A limitation of our study is that the findings reported were made through an analysis of existing cancer genomics data. Future experimental studies can be performed to confirm the underlying mechanism linking mutant *ERCC2* to altered mutational distribution. Although our research has already been conducted on the largest BLCA whole-genome dataset to date (n = 382), the number of whole cancer genomes with these mutations is still relatively small due to the modest frequency of ERCC2 mutations (∼10%). In the future, larger cohorts may help identify weaker associations between genomics features and their mutation burden. Another limitation of our study is that the retrospective analysis of existing data means we do not have broader access to the patient's clinical parameters. A prospective study or a clinical trial will help validate the clinical utility of *ERCC2* mutation status in predicting BLCA patient prognosis and fully evaluate its impact on chemotherapy.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Somatic mutations and simulation
  - Calculation of local mutation densities and generation of mutation profiles across genomic sites
  - Mutation trinucleotide frequency calculations
  - Genomic annotations and data binning
  - CTCF ChIP coverage at CBS
  - Generating coverage profiles across genomic regions for ChIP-seq data
  - Uracil sequencing data
  - Support machine vector model
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100627.

ACKNOWLEDGMENTS

This work was enabled by access to data and findings generated by the 100,000 Genomes Project, under the Pan-Cancer Genomics England Clinical Interpretation Partnership (GeCIP), project ID 752. The project was supported by the Research Grants Council, Hong Kong SAR, China (17108823, 17100920, and R7022-20), to J.W.H.W. and the Centre for Oncology and Immunology under the Health@InnoHK Initiative funded by the Innovation and Technology Commission, the Government of Hong Kong SAR, China.

AUTHOR CONTRIBUTIONS

J.A.B. analyzed data, prepared figures, and wrote the manuscript; H.Y., H.F., N.C.Y., X.Z., and Y.T.W. analyzed and interpreted data; T.O., M.W.W.-B., and N.A.B. contributed data; S.W. contributed data and jointly supervised the research; and J.W.H.W. conceived and designed the study, analyzed data, revised the manuscript, and supervised the research.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 12, 2024
Revised: May 17, 2024
Accepted: July 10, 2024
Published: August 2, 2024

REFERENCES

1. Singh, A., Compe, E., Le May, N., and Egly, J.M. (2015). TFIIH subunit alterations causing xeroderma pigmentosum and trichothiodystrophy specifically disturb several steps during transcription. Am. J. Hum. Genet. *96*, 194–207. https://doi.org/10.1016/j.ajhg.2014.12.012.
2. Yan, C., Dodd, T., He, Y., Tainer, J.A., Tsutakawa, S.E., and Ivanov, I. (2019). Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. Nat. Struct. Mol. Biol. *26*, 397–406. https://doi.org/10.1038/s41594-019-0220-3.
3. Tsutakawa, S.E., Bacolla, A., Katsonis, P., Bralić, A., Hamdan, S.M., Lichtarge, O., Tainer, J.A., and Tsai, C.L. (2021). Decoding Cancer Variants of Unknown Significance for Helicase-Nuclease-RPA Complexes Orchestrating DNA Repair During Transcription and Replication. Front. Mol. Biosci. *8*, 791792. https://doi.org/10.3389/fmolb.2021.791792.
4. Van Allen, E.M., Mouw, K.W., Kim, P., Iyer, G., Wagle, N., Al-Ahmadie, H., Zhu, C., Ostrovnaya, I., Kryukov, G.V., O'Connor, K.W., et al. (2014). Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. Cancer Discov. *4*, 1140–1153. https://doi.org/10.1158/2159-8290.Cd-14-0623.
5. Comprehensive molecular characterization of urothelial bladder carcinoma. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322. https://doi.org/10.1038/nature12965.
6. Li, Q., Damish, A.W., Frazier, Z., Liu, D., Reznichenko, E., Kamburov, A., Bell, A., Zhao, H., Jordan, E.J., Gao, S.P., et al. (2019). ERCC2 Helicase Domain Mutations Confer Nucleotide Excision Repair Deficiency and Drive Cisplatin Sensitivity in Muscle-Invasive Bladder Cancer. Clinical cancer research. an official journal of the American Association for Cancer Research *25*, 977–988. https://doi.org/10.1158/1078-0432.Ccr-18-1001.
7. Börcsök, J., Sztupinszki, Z., Bekele, R., Gao, S.P., Diossy, M., Samant, A.S., Dillon, K.M., Tisza, V., Spisák, S., Rusz, O., et al. (2021). Identification of a Synthetic Lethal Relationship between Nucleotide Excision Repair Deficiency and Irofulven Sensitivity in Urothelial Cancer. Clin. Cancer Res. *27*, 2011–2022. https://doi.org/10.1158/1078-0432.Ccr-20-3316.
8. Liu, D., Plimack, E.R., Hoffman-Censits, J., Garraway, L.A., Bellmunt, J., Van Allen, E., and Rosenberg, J.E. (2016). Clinical Validation of Chemotherapy Response Biomarker ERCC2 in Muscle-Invasive Urothelial Bladder Carcinoma. JAMA Oncol. *2*, 1094–1096. https://doi.org/10.1001/jamaoncol.2016.1056.
9. Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606. https://doi.org/10.1038/ng.3557.
10. Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is

widespread in human cancers. Nat. Genet. *45*, 970–976. https://doi.org/10.1038/ng.2702.

11. Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504–507. https://doi.org/10.1038/nature11273.

12. Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81–84. https://doi.org/10.1038/nature14173.

13. Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat. Genet. *49*, 1684–1692. https://doi.org/10.1038/ng.3991.

14. Perera, D., Poulos, R.C., Shah, A., Beck, D., Pimanda, J.E., and Wong, J.W. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature *532*, 259–263. https://doi.org/10.1038/nature17437.

15. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature *532*, 264–267. https://doi.org/10.1038/nature17661.

16. Poulos, R.C., Olivier, J., and Wong, J.W.H. (2017). The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. Nucleic Acids Res. *45*, 7786–7795. https://doi.org/10.1093/nar/gkx463.

17. Clinton, T.N., Chen, Z., Wise, H., Lenis, A.T., Chavan, S., Donoghue, M.T.A., Almassi, N., Chu, C.E., Dason, S., Rao, P., et al. (2022). Genomic heterogeneity as a barrier to precision oncology in urothelial cancer. Cell Rep. *41*, 111859. https://doi.org/10.1016/j.celrep.2022.111859.

18. Kinnersley, B., Sud, A., Everall, A., Cornish, A.J., Chubb, D., Culliford, R., Gruber, A.J., Larkeryd, A., Mitsopoulos, C., Wedge, D., and Houlston, R. (2024). Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. Nat. Genet., 1–10. https://doi.org/10.1038/s41588-024-01785-9.

19. Lada, A.G., Kliver, S.F., Dhar, A., Polev, D.E., Masharsky, A.E., Rogozin, I.B., and Pavlov, Y.I. (2015). Disruption of Transcriptional Coactivator Sub1 Leads to Genome-Wide Re-distribution of Clustered Mutations Induced by APOBEC in Active Yeast Genes. PLoS Genet. *11*, e1005217. https://doi.org/10.1371/journal.pgen.1005217.

20. Chan, K., Sterling, J.F., Roberts, S.A., Bhagwat, A.S., Resnick, M.A., and Gordenin, D.A. (2012). Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. PLoS Genet. *8*, e1003149. https://doi.org/10.1371/journal.pgen.1003149.

21. Petljak, M., Dananberg, A., Chu, K., Bergstrom, E.N., Striepen, J., von Morgen, P., Chen, Y., Shah, H., Sale, J.E., Alexandrov, L.B., et al. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. Nature *607*, 799–807. https://doi.org/10.1038/s41586-022-04972-y.

22. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.

23. Wu, S., Ou, T., Xing, N., Lu, J., Wan, S., Wang, C., Zhang, X., Yang, F., Huang, Y., and Cai, Z. (2019). Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications as angiogenesis-related drivers in bladder cancer. Nat. Commun. *10*, 720. https://doi.org/10.1038/s41467-019-08576-5.

24. Shen, P., Jing, Y., Zhang, R., Cai, M.C., Ma, P., Chen, H., and Zhuang, G. (2018). Comprehensive genomic profiling of neuroendocrine bladder cancer pinpoints molecular origin and potential therapeutics. Oncogene *37*, 3039–3044. https://doi.org/10.1038/s41388-018-0192-5.

25. Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong, J.W.H. (2016). Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. Cell Rep. *17*, 2865–2872. https://doi.org/10.1016/j.celrep.2016.11.055.

26. Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821. https://doi.org/10.1038/ng.3335.

27. Guo, Y.A., Chang, M.M., Huang, W., Ooi, W.F., Xing, M., Tan, P., and Skanderup, A.J. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat. Commun. *9*, 1520. https://doi.org/10.1038/s41467-018-03828-2.

28. Kaiser, V.B., Taylor, M.S., and Semple, C.A. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet. *12*, e1006207. https://doi.org/10.1371/journal.pgen.1006207.

29. Gray, L.T., Vallur, A.C., Eddy, J., and Maizels, N. (2014). G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. Nat. Chem. Biol. *10*, 313–318. https://doi.org/10.1038/nchembio.1475.

30. Hu, J., Lieb, J.D., Sancar, A., and Adar, S. (2016). Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proc. Natl. Acad. Sci. USA *113*, 11507–11512. https://doi.org/10.1073/pnas.1614430113.

31. Jiang, L., Yin, J., Qian, M., Rong, S., Zhang, S., Chen, K., Zhao, C., Tan, Y., Guo, J., Chen, H., et al. (2022). UdgX-Mediated Uracil Sequencing at Single-Nucleotide Resolution. J. Am. Chem. Soc. *144*, 1323–1331. https://doi.org/10.1021/jacs.1c11269.

32. Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. *23*, 517–525. https://doi.org/10.1038/nm.4292.

33. Telli, M.L., Timms, K.M., Reid, J., Hennessy, B., Mills, G.B., Jensen, K.C., Szallasi, Z., Barry, W.T., Winer, E.P., Tung, N.M., et al. (2016). Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. Clin. Cancer Res. *22*, 3764–3773. https://doi.org/10.1158/1078-0432.CCR-15-2477.

34. Polak, P., Lawrence, M.S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R.E., Garraway, L.A., Mirkin, S., Getz, G., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2014). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat. Biotechnol. *32*, 71–75. https://doi.org/10.1038/nbt.2778.

35. Chatzinikolaou, G., Apostolou, Z., Aid-Pavlidis, T., Ioannidou, A., Karakasilioti, I., Papadopoulos, G.L., Aivaliotis, M., Tsekrekou, M., Strouboulis, J., Kosteas, T., and Garinis, G.A. (2017). ERCC1-XPF cooperates with CTCF and cohesin to facilitate the developmental silencing of imprinted genes. Nat. Cell Biol. *19*, 421–432. https://doi.org/10.1038/ncb3499.

36. Lake, R.J., Boetefuer, E.L., Won, K.J., and Fan, H.Y. (2016). The CSB chromatin remodeler and CTCF architectural protein cooperate in response to oxidative stress. Nucleic Acids Res. *44*, 2125–2135. https://doi.org/10.1093/nar/gkv1219.

37. Christensen, S., Van der Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens, J.W.M., Yaspo, M.L., Priestley, P., Kuijk, E., Cuppen, E., and Van Hoeck, A. (2019). 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat. Commun. *10*, 4571. https://doi.org/10.1038/s41467-019-12594-8.

38. An, Q., Robins, P., Lindahl, T., and Barnes, D.E. (2005). C–> T mutagenesis and gamma-radiation sensitivity due to deficiency in the Smug1 and Ung DNA glycosylases. EMBO J. *24*, 2205–2213. https://doi.org/10.1038/sj.emboj.7600689.

39. Kim, N., and Jinks-Robertson, S. (2009). dUTP incorporation into genomic DNA is linked to transcription in yeast. Nature *459*, 1150–1153. https://doi.org/10.1038/nature08033.

40. Owiti, N., Wei, S., Bhagwat, A.S., and Kim, N. (2018). Unscheduled DNA synthesis leads to elevated uracil residues at highly transcribed genomic loci in Saccharomyces cerevisiae. PLoS Genet. *14*, e1007516. https://doi.org/10.1371/journal.pgen.1007516.

41. Kavli, B., Sundheim, O., Akbari, M., Otterlei, M., Nilsen, H., Skorpen, F., Aas, P.A., Hagen, L., Krokan, H.E., and Slupphaug, G. (2002). hUNG2 is the major repair enzyme for removal of uracil from U: A matches, U: G mismatches, and U in single-stranded DNA, with hSMUG1 as a broad specificity backup. J. Biol. Chem. *277*, 39926–39936. https://doi.org/10.1074/jbc.M207107200.

42. Fortin, J.P., and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. *16*, 180. https://doi.org/10.1186/s13059-015-0741-y.

43. Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., Furney, S.J., and Lopez-Bigas, N. (2010). IntOGen: integration and data mining of multidimensional oncogenomic data. Nat. Methods *7*, 92–93. https://doi.org/10.1038/nmeth0210-92.

44. Bergstrom, E.N., Barnes, M., Martincorena, I., and Alexandrov, L.B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. BMC Bioinf. *21*, 438. https://doi.org/10.1186/s12859-020-03772-3.

45. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

46. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31. https://doi.org/10.1186/s13059-016-0893-4.

47. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801. https://doi.org/10.1093/nar/gkx1081.

48. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. https://doi.org/10.1093/nar/gku365.

49. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
HEPG2 Replication Time	UCSC Table Browser	wgEncodeEH002244
A/B compartments	Fortin et al.[42]	Supplementary data
Mutations from APOBEC expressing cell lines	Petljak et al.[21]	Table S8 from source
Bladder DNase I hypersensitivity	ENCODE	ENCSR813CKU
Bladder H3K4Me ChIP	ENCODE	ENCSR632OWD
Bladder H3K27Ac ChIP	ENCODE	ENCSR054BKO
CTCF-cohesin binding site	Katainen et al.[26]	Supplementary data
Kidney CTCF ChIP	ENCODE	ENCSR000DMC
XPD and XPB ChIP	GEO	GSE44849
TFIIH XR-seq	GEO	GSE82213
Ucaps-seq	SRA	PRJNA728500
MSK Bladder Cancer Cohort	cBioportal	paired_bladder_2022
Genomics England Bladder Cancer Cohort	GE Research Environment	N/A
PCAWG Cohort Mutations	ICGC Data Portal	consensus_snv_indel
Cohort of Neuroendocrine Bladder Mutations	SRA	PRJNA399789
Other cohort of Urothelial Bladder Carcinoma	EGA	EGAS00001003388
Software and algorithms		
R (version 4.3.2)	The R Foundation	https://www.r-project.org/
Rstudio (version 2023.12.0)	Posit Software	https://posit.co/products/open-source/rstudio/
survminer	GitHub	https://github.com/kassambara/survminer
finalfit	GitHub	https://github.com/ewenharrison/finalfit
forestploter	GitHub	https://github.com/adayim/forestploter
SigProfilerSimilator	Bergstrom et al.[44]	https://github.com/AlexandrovLab/SigProfilerSimulator
Bedtools	Quinlan et al.[45]	https://github.com/arq5x/bedtools2
DeconstructSigs	Rosenthal et al.[46]	https://github.com/raerose01/deconstructSigs
Deeptools	Ramirez et al.[47]	https://github.com/deeptools/deepTools
Uracil analysis	Jiang et al.[31]	https://github.com/Jyyin333/Ucaps-seq
Picard	Broad Institute	http://broadinstitute.github.io/picard
E1071	GitHub	https://github.com/cran/e1071

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jason W.H. Wong (jwhwong@hku.hk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

The original code has been deposited at https://github.com/jasonwong-lab/ERCC2 (https://doi.org/10.5281/zenodo.12676717) and is publicly available as of the date of publication. Additional code is available in the Genomics England Research Environment at/re_gecip/cancer_pan/jbarbour/scripts/. The link to become a member of the Genomics England research network and obtain access can be found at https://www.genomicsengland.co.uk/research/academic/join-gecip.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

For the analysis of the independent prognostic significance of *ERCC2* mutations, somatic mutation calls and patient clinical characteristics were obtained for 1,244 from the Memorial Sloan Kettering (MSK) BLCA cohort.[17] Somatic mutation calls were obtained from the cBioportal repository. *ERCC2* mutations were classified as recurrent missense hotspot, other missense or nonsense. For the definition of recurrent missense hotspot mutations, the AACR-GENIE database (n = 137,401) was used to find recurrent *ERCC2* mutations, with recurrence defined as a protein-altering mutation in *ERCC2* found in at least one BLCA sample plus one other sample of any cancer type. For patients with multiple samples, primary samples with earlier specimen stage were used. Patients without complete age, mutation count, and specimen stage information were excluded from the analysis. In total, 113 samples were classified as *ERCC2* mutant and 886 as *ERCC2* WT.

For the analysis of genome-wide mutation distribution, somatic mutation calls from 392 Genomics England (GE)[18] whole-genome sequenced (WGS) BLCA samples were accessed directly from the GE research environment. Of the 392 GE BLCA samples, 39 contained recurrent *ERCC2* mutations, were classified as *ERCC2* mutant and 343 samples that had no protein-altering mutations in *ERCC2* were classified as WT. A further ten samples had protein-altering mutations in *ERCC2* that were not recurrent. These ten samples were excluded from analysis as we could not confidently assign them as either WT or *ERCC2* mutant (Table S1).

For both MSK and GE cohorts, to analyze the proportion of oncogene protein-altering mutations in WT and *ERCC2* mutant groups, we counted the number of samples in each group with either missense, nonsense or frameshift mutations in genes from the IntOGen database.[43]

For SBS17 and SBS7 cancers, we used somatic mutations from GE esophageal adenocarcinoma (ESAD) and GE melanoma (MELA), respectively. This resulted in 3,692,032 mutations from 106 samples for ESAD and 363,284,245 mutations from 337 samples for MELA.

For analysis of APOBEC mutations in cell lines, data was accessed from.[21] This dataset included somatic mutation data from single-cell clones and included the following cell lines with endogenous APOBEC mutational signatures – BC-1, BT-474, HT-1376, JSC-1 and MDA-MB-453. BT-474 and MDA-MB-453 cells also had mutations from clones that were either WT and *UNG* (encoding UDG2) knockout by CRISPR. We pooled daughter mutations together by cell line.

We additionally utilised somatic mutations from Pan-cancer analysis of whole genomes (PCAWG) BLCA and liver cancer,[22] a study of WGS urothelial bladder carcinomas[23] and a study of WGS neuroendocrine bladder cancer.[24] The PCAWG BLCA study had 4 *ERCC2* mutant and 19 WT samples. The urothelial bladder carcinoma study had 2 *ERCC2* mutant and 63 WT. The neuroendocrine BLCA had 1 *ERCC2* mutant and 5 WT samples. These 3 studies of BLCA were combined, giving 7 *ERCC2* mutant and 87 WT. For liver cancer samples from PCAWG, there were 3 *ERCC2* mutant samples and 302 WT.

METHOD DETAILS

Somatic mutations and simulation

For GE cohorts, SNV calls for hg38 were obtained directly from the GE research environment (RE) and hg38 annotations were used for these analyses. For BLCA, SNVs that were C > D (D represents A, G or T) at TCN context were defined as APOBEC, whilst all SNVs not in this context were defined as ‘Other’ (Other). Other mutations were assigned as Other [C>T]pG and Other non-[C>T]pG. APOBEC mutations were separated into SBS2 or SBS13 like if they were T[C>T]N or T[C>R]N, respectively. Simulations were used to establish the chance of genomic positions being mutated based on sequence context and mutation burden using SigProfilerSimulator.[44] For BLCA and ESAD, 100 simulations were performed and were merged and divided by 100 giving what we refer to as ‘expected’. For MELA mutations, 10 simulations were used for the expected calculations due to memory constraints. For mutations from APOBEC expressing cell lines and other studies, including PCAWG, hg19 mutation calls were used and hg19 annotations were used to generate those figures.

Calculation of local mutation densities and generation of mutation profiles across genomic sites

To calculate mutation densities at specific genomic regions, we counted the number of actual mutations (observed) and simulated mutations overlapping these regions using the tool ‘intersectBed’.[45] Mutations of 100 or 10 simulations were merged for analysis and then divided by 100 or 10 to give an “expected” value, and then local mutation density was expressed as the ratio of observed to expected mutations for profile plots and as log2(obs/exp) for mutation densities in all regions expect CBS and flank. For mutation densities for CBS and flank, due to the low number of expected mutations falling in these regions, we calculated as log2 (obs+1/exp+1).

Genome-wide distribution of mutations was performed by calculating mutation densities as described above for one megabase (mb) window of the human genome. Mutation densities for chromosome 1 were plotted positionally, and then principal component analysis (PCA) was performed genome wide in R using prcomp() function with scaling and centering. To perform statistics on local mutation densities of bins based on genomic coverage, we calculated the mean coverage of each bin and performed linear regression between mutation densities and coverage for each sample, displaying the mean and standard deviation and regression line on the graph. To generate mutation density profiles across regions, windows were generated within, upstream or downstream, with each region separated according to the number of bins and number of bases flank specified. Where regions contain sites of varied lengths e.g., gene bodies, the number of windows for each site was fixed, therefore changing the number of bases per window in the region. Mutation densities were then calculated in each of the windows as described above.

Mutation trinucleotide frequency calculations

Mutation trinucleotide frequencies were calculated using DeconstructSigs.[46] For frequencies across the whole genome, "genome" normalisation was used. For frequencies on specific regions, such as CBS motifs, the trinucleotide composition was calculated using grep scripts, and then these were input into DeconstructSigs with "manual" normalisation.

Genomic annotations and data binning

Gene expression data was taken from the GTEx portal and the top half of expressed genes were defined as "expressed" in bladder. Genes with 0 counts in bladder were defined as "silent". Annotations of genic regions including, 5′ untranslated region (UTR), 3′ UTR, exons and introns were accessed from UCSC table browser for hg19. Intergenic regions for hg19 were defined as parts of the genome without overlap of any of these regions. Hg19 coverage and narrow peaks data for human bladder tissue DNase-seq experiments were accessed from ENCODE[47] (ENCSR813CKU) as bigWig and bed file, respectively. TFIIH XR-seq data was accessed as a bigwig from GEO under accession GSE82213.[30] 1 kb windows of hg19 were generated and then filtered for blacklisted and low-coverage regions of the genome. To divide the genome into bins based on coverage of different genomic assays, including DNase-seq, replication time, TFIIH XR-seq ChIP-seq, the mean bedgraph signal from genomic assays was calculated for each of the 1kb filtered genomic windows using bedtools map.[45] For mutation density calculations, these filtered 1 kb windows were then divided into quintiles based on coverage from lowest signal (bin 1) to highest signal (bin 5).

Bladder DHS peaks were overlapped with other DHS marks to generate annotations for bladder DNase hypersensitive regions (DHS) as follows. Promoters were defined by overlap with bladder H3K4Me3 ChIP-seq peaks from ENCODE (ENCSR632OWD), and then gene start sites to get promoters. Bladder DHS peaks were overlapped with high quality, experimentally determined CBS accessed from supplementary materials of ref. 26 to generate CBS annotations. Later analysis of CBS uses these high quality CBS annotations[26] without overlapping with bladder DHS. Finally, enhancers were defined as the center of bladder H3K27Ac ChIP-seq peaks from ENCODE (ENCSR054BKO) that overlapped bladder DHS peaks. Chromatin A/B compartments for the bladder were taken from supplementary files of.[42] For later analysis on CBS, all 31252 CBS were used.[26] The above annotations were converted to hg38 using "liftOver". CBS were further annotated for certain analyses. CBS were defined as either genic or intergenic based on overlap with canonical genes from the UCSC table browser. CBS were also identified as either "mutated" or "not-mutated" based on if the site had the presence or absence of an Other mutation from *ERCC2* mutant GE BLCA samples, respectively. For analysis of CTCF occupancy, CBS were divided into quartiles using CTCF ChIP data acquired from kidney tissue from ENCODE (ENCSR000DMC). Briefly, a CTCF over input fold change bigwig file was accessed, and the CTCF ChIP-seq coverage was mapped to CBS using bedtools map. CBS were then arranged into quartiles based on ascending fold change where Q1 represents low CTCF coverage and Q4 represents high. CBS were annotated as uracil containing or not uracil containing based on Ucaps Seq data (see below).

CTCF ChIP coverage at CBS

Fold change over control bigwig file for kidney CTCF ChIP was directly accessed from ENCODE (ENCSR000DMC). This was converted to bedgraph using ucsc tools and then the average coverage was mapped to CBS that were either mutated or not mutated or CBS with or without uracil.

Generating coverage profiles across genomic regions for ChIP-seq data

BigWig files (hg19) for XPD and XPB ChIP-seq and input were accessed from gene expression omnibus (GEO) under accession GSE44849, which was previously published.[29] Deeptools "bigwigCompare"[48] was used to generate log2 ratio bigwig files of the ChIP compared with input, with a pseudocount of 1. The centerpoint of hg19 genic and intergenic CBS was retrieved and deeptools "computeMatrix" was used to calculate the signal around 1000 bases organised into 200 bins both upstream and downstream of the centerpoint. log2 ratio ChIP/input signals were averaged and individual data points were displayed as heatmaps using pheatmap in R.

Uracil sequencing data

Uracil sequencing data (Ucaps-seq) previously published[31] was accessed from European Nucleotide Archive under accession PRJNA728500. Data was processed as the authors described. Briefly, fastq files were trimmed with bbduk and aligned to hg19 using bwa,[49] then duplicates were removed using picard (http://broadinstitute.github.io/picard). Single base locations of uracil were located using scripts "fetch_dU_by_chrom.py" from https://github.com/Jyyin333/Ucaps-seq. If the reference base matched a T/A or a C/G it was considered to be from incorporation or deamination, respectively. Uracil sequencing data profiles were drawn around hg19 CBS using coverageBed.[45] Bedtools slop and fastaFromBed were used to retrieve the trinucleotide sequence context of uracil. The number of uracil in the input file was subtracted from the experiment file for specific regions.

Support machine vector model

Results of somatic mutation densities from GE for the following genomic regions: CBS motif, 1000 bp flanking CBS motif, coding exons, introns, 3′UTR, 5′UTR, intergenic, open and closed chromatin and 5 regions of the genome binned by replication time, DNase hypersensitivity, XPD ChIP-seq coverage and TFIIH XR-seq coverage was used as the input to train a support vector machine (SVM) for classifying whether a sample was a driver or passenger mutation. The svm function from the e1071 R package was used. The SVM model was first evaluated using leave-one-out-cross validation using the GE cohort (samples with well-defined *ERCC2* mutation status). We further trained the SVM model using all of these GE samples and tested this on the independent TCGA cohort. Finally, we used the same model to evaluate the *ERCC2* mutation status of the GE samples with undefined *ERCC2* mutation status. The above method was applied to generate a model with the input simply being the trinucleotide context of SNVs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Survival and hazard ratio analysis (univariable and multivariable) were performed in Rstudio (Version 2023.12.0) using R (version 4.3.2). The survival analysis was conducted utilising the open-source R package survminer (version 0.4.9), and group differences were assessed by the log rank test. The hazard ratios were calculated using the proportional hazard regression model in the finalfit (version 1.0.7) and then formatted by forestploter (version 1.1.1). Other statistical tests including Student’s t test and Mann-Whitney test were performed using R or Graphpad PRISM. A *p*-value of less than or equal to 0.05 was considered statistically significant.