XPD protects CTCF-Cohesin binding sites from somatic mutagenesis

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

[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China

[2]Urology Institute of Shenzhen University, The Third Affiliated Hospital of Shenzhen University, Shenzhen University, Shenzhen, China

[3]Centre for Drug Repurposing and Medicines Research, University of Newcastle, NSW, Australia

[4]Hunter Medical Research Institute, Newcastle, NSW, Australia

[5]Adult Cancer Program, Lowy Cancer Research Centre, UNSW Sydney, NSW, Australia

[6]Centre for PanorOmic Sciences, The University of Hong Kong, Pokfulam, Hong Kong Special Administrative Region, China

*correspondence should be addressed to:

Jason W. H. Wong, jwhwong@hku.hk

Song Wu, wusong@szu.edu.cn

Abstract

Xeroderma pigmentosum group D (XPD) is a DNA helicase involved in transcription initiation and nucleotide excision repair. Missense mutations in XPD are putative drivers in bladder cancer (BLCA) and are associated with a specific single base substitution mutational signature. However, the impact of XPD on the genome-wide distribution of somatic mutations remains unexplored. We analysed somatic mutation distribution in whole-genome sequenced (WGS) BLCA samples with (XPD mutant) and without XPD mutations (WT). XPD genotype had a large impact on the distribution of somatic mutations. XPD mutant samples had increased mutation density at open chromatin, including striking mutation hotspots at CTCF-cohesin binding sites (CBS). We validate these findings in additional WGS cohorts and BLCA exomes. Analysis of XPD occupancy and CBS hotspot mutations in other cancer types suggest that XPD protects CBS from DNA damage. Our study implicates XPD in genomic integrity maintenance at topologically-associating domain boundaries marked by CTCF-cohesin binding.

Introduction

Xeroderma pigmentosum group D (XPD), encoded by *ERCC2*, is a 5'-3' ATP-dependent DNA helicase that is a component of the Transcription Factor II H (TFIIH) protein complex. TFIIH plays important roles in transcription initiation through its interaction with RNA polymerase II (POLR2A) and nucleotide excision repair (NER) when recruited to damaged lesions [1]. Compound heterozygous mutations in XPD can cause the genetic disorders xeroderma pigmentosum and trichothiodystrophy which typically present with UV light sensitivity due to deficiencies in NER function [2]. Additionally, somatic missense mutations in XPD are putative drivers in BLCA with ~ 12% of predominantly Caucasian cohorts of BLCA samples harbouring these alterations [3, 4]. XPD mutant BLCA are sensitive to cisplatin therapy, indicating a reduced capacity for repair of cisplatin adduct DNA lesions which implies a deficiency in NER of these samples [3, 5].

Somatic mutation density in cancer forms specific patterns across the genome in terms of the mutation profiles and regional mutation densities. The type of single nucleotide variant (SNV, referred to as mutation) in the trinucleotide context forms specific single base substitution (SBS) mutational spectra or signatures which reflect the mutational process of the sample [6, 7]. Genomic mutation density is highly varied across the genome, correlating strongly with various epigenetic marks such as chromatin accessibility [8], histone modifications [9, 10], transcription factor binding [11, 12] and cytosine methylation [13]. Regions of the genome with exceptionally high mutation densities are considered 'mutational hotspots'. One such hotspot is CCCTC-binding factor (CTCF)-cohesin binding sites (CBS) of which there have been several reports of strongly elevated somatic mutation densities [14-19]. CTCF is a DNA binding protein that acts as a transcriptional repressor when bound to DNA alone and as an architectural protein when

CTCF proteins bound at distal sites dimerise and interact with the cohesin complex to form DNA loops [20]. Interestingly, CBS hotspots have previously been linked to specific mutational signatures cosmic SBS7 [17] and SBS17 [14, 18].

XPD mutant BLCA has previously been associated with the enrichment of mutational signature, SBS5 [21], however, how mutant XPD causes this mutational signature remains unknown. To gain an insight into the XPD mutant driven mutational process we compared with mutation distribution of XPD wild-type and mutant bladder cancers across a range of genetic and epigenetic features. Beyond the known role of XPD in transcription coupled-NER, our findings point to XPD protecting open chromatin from mutations. In particular, we observed strong mutational hotspots at CBS in XPD mutants, suggesting a previously unknown role of XPD in topologically associating domain (TAD) boundary maintenance.

Results

Differential Contribution of APOBEC Associated and Other Mutations in XPD Mutant and Wild Type Bladder Cancer

XPD mutations have been linked to a specific mutational signature in BLCA [21], but the genome-wide distribution of mutations associated with XPD mutants is unknown. To investigate this, we utilised the TCGA cohort of WGS BLCA and characterised samples that harboured putative XPD driver mutations. Out of a total of 23 samples, 4 were characterised as XPD mutant. Previous studies have found the presence of strong APOBEC mutational signatures in many BLCA samples [22] and signature 5 (SBS5) specifically in XPD mutant BLCA samples, but analysis was restricted to exomes [21]. 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 XPD mutant and non-mutant BLCA. We hypothesised that if we separate the APOBEC related mutations from Other mutations in the sample, we can better delineate the XPD mutational processes. This is feasible due to the highly specific nature of APOBEC mutations [22]. To this end, we assigned C>D at TCN as APOBEC and all else as Other. XPD mutant samples have a tendency for more Other mutations than WT (Figure 1A and Supplementary table 1). After separating APOBEC from Other mutations (Figure 1B, Figure S1A), we found that, as expected, the cosine similarity of APOBEC mutations was more similar to SBS2 and SBS13 than for all mutations (Figure 1C), while other mutations were more similar to SBS5 than all mutations (Figure 1C). This demonstrates that we can largely distinguish these mutational processes using this method. We next investigated the distribution of APOBEC and Other mutations across the genome. We found that the genome-wide distribution of Other mutations but not APOBEC mutations differ between WT and XPD mutant

samples (Figure 1D). This suggests that XPD specific mutational processes are unique and validates the concept of separating APOBEC mutations from analysis to uncover these genomic patterns. We did not observe any significant genome-wide differences in indel and structural variant counts between XPD WT and mutant samples (Figure S1B).

XPD Mutant Samples Display Altered Genomic Distribution of Other Mutations

Somatic mutations in human cancer are unevenly distributed across the genome, with mutations in most mismatch repair proficient cancers showing reduced mutation burden at early replicating regions of the genome which are associated with open chromatin and expressed genes [8]. However, for certain mutations processes such as APOBEC-induced mutations, this trend does not apply [23]. To explore mutation distribution further, mutation densities for APOBEC and Other mutations were calculated with respect to gene bodies and replication time in XPD mutant and WT BLCA. We found that the distribution of Other mutations is significantly higher in all genic regions and lower in intergenic regions in XPD mutant samples compared with WT, which is particularly pronounced for the 5'UTR ($q=0.007171$, Student's t-test with multiple testing correction, Figure 2A). However, there were no significant differences between XPD mutant and WT samples for APOBEC mutations ($q>0.33$, Student's t-test with multiple testing correction, Figure 2A). We found an increase in the burden of APOBEC related mutations in 5'UTR relative to what is expected by chance (observed-expected ratio > 1) in both XPD mutant and WT groups (Figure 2A). This is consistent with previous findings that APOBEC causes mutation clusters around the start of active genes [24]. A linear regression between mutation densities and replication time showed that WT samples had a slope of -0.02601 compared with -0.005661 for mutant (Figure 2B), with significantly decreased and increased burdens of mutations in mutant

samples compared with WT in late and early replicating regions respectively (q=0.000197 and q=0.000015, Student's t-test with multiple testing correction, Figure S2A). APOBEC mutations remained largely unchanged across the replication time landscape as consistent with previous literature [23], and this pattern is not affected by XPD mutations (Figure 2B, Figure S2A). 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 XPD related mutagenesis. We next examined the effect that transcriptional activity has on mutagenesis in the BLCA genomes. We find that XPD mutant samples have increased genic mutation burden for Other but not APOBEC mutations, compared to WT specifically at expressed genes (Figure 2C). This was particularly pronounced immediately before the transcriptional start site (TSS) (Figure 2C and S2C), which is consistent with our results from Figure 2A.

To look more generally at active and inactive chromatin genome-wide, we next used DNase hypersensitivity (DHS) to compare the burden of APOBEC and Other mutations between XPD mutant and WT samples. Interestingly, we find that Other mutations in XPD mutant samples tend to accumulate in DHS regions (Figure 2D), with significantly more mutations in XPD mutants compared with WT in the most DHS regions (q=0.000361, Student's t-test with multiple testing correction, Figure S2C). We also found significantly increased and decreased burden of Other mutations in XPD mutant samples in open compartments and closed compartments respectively compared with WT (q=0.000067, q=0.000067, Student's t-test with multiple testing correction, Figure 2E). These observations remained the same even when restricting the analysis to intergenic and non-CBS regions (Figure S2D). This provides strong evidence that XPD mutagenesis is enhanced at accessible chromatin.

XPD Mutant Cancers Display Strong Mutation Hotspots in CTCF-Cohesin Binding Sites

DHS are associated with *cis*-regulatory elements, including promoters, enhancers and CTCF-cohesin binding sites (CBS). As Other mutations in XPD mutants showed increased mutation density at DHS regions, we examined these elements individually and found striking mutation hotpots in CBS but not promoters or enhancers (Figure 3A).

While somatic mutation hotspots in CBS have previously been reported in UV associated skin cancers [17] and SBS17 associated gastrointestinal cancers [14, 16, 18] it is a striking and novel observation that XPD associated mutational signatures in BLCA also have CBS mutation hotspots. Therefore, we sought additional samples to validate these findings. We accessed 2 Chinese cohorts of BLCA [25, 26] and found an additional 3 samples with XPD mutations and found 3 additional XPD mutant liver cancer samples from PCAWG. Combined with the TCGA BLCA, there were 10 XPD mutant samples for these corresponding cohorts. We plotted the density enrichment at CBS and flank and found XPD mutant samples observed elevated mutation densities in CBS in XPD mutant compared to WT (Figure 3B, Figure S3A). We also observed increased mutation densities in flanking regions in XPD mutant samples compared with WT but this is likely reflects generally greater chromatin accessibility of the CBS flank and, in any case is substantially lower compared with the CBS itself (Figure 3B). A proportion of CBS sites lies within the covered regions of exome capture (n = 1049), so we further generated a contingency table of mutations from TCGA exome sequenced BLCA samples and found >4-fold enrichment for mutations at CBS in XPD mutant samples compared with WT samples ($p<0.0001$, Fisher's exact test, Figure 3C,). Collectively, these results provide strong evidence that XPD mutant cancer displays hotspots in CBS.

Previous reports of CBS hotspots had found specific mutational patterns and signatures across the CBS motif [14, 17]. We observed that the CBS mutations in XPD mutants also have a similar distribution compared with CBS hotspots found in esophageal adenocarcinoma (ESAD) but is different to melanoma (MELA) (Figure 3D, Figure S3B, C). 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 3E). This is similar to gastrointestinal cancers with SBS17 where predominantly T>G and T>C mutations accumulate at CBS [14, 18].

Effect of XPD Presence on Genomic Mutation Distribution

We next wanted to explore how XPD itself affects the development of mutations in WT and mutant cancers. Using previously published XPD ChIP-seq data [27], we examined mutagenesis in genomic regions with respect to XPD coverage. We found a strong enrichment of XPD at CBS, and as XPB is also enriched at CBS (Figure 4A), implying that these proteins are co-bound to CBS as part of the TFIIH complex. We next examined the general relationship between XPD binding and mutation burden in XPD mutant and WT BLCA. We found that WT BLCA had a distinct pattern of increased somatic mutation densities at low XPD coverage regions and decreased mutation densities at high XPD coverage regions, a trend which is reduced in XPD mutant samples with slopes of -0.3261 and -0.09264, respectively (Figure 4B and S4A). This suggests that the WT XPD protein likely plays a protective role where it is bound in the genome. We found that the mutational spectra were highly similar across XPD high and low binding regions in XPD mutant and WT cancers (Figure 4C). Therefore, the amount of XPD coverage in the genome affects mutation densities, not mutation types. If mutant XPD was actively causing

damage or repair errors, we would expect a specific mutational signature to be present at high XPD coverage regions of the genome that should be absent in low XPD coverage regions. A change in mutation density but not mutation types with XPD coverage is consistent with a loss of dependence on DNA repair.

These observations led us to hypothesise that WT XPD has a role in protecting genomic integrity either through DNA repair or by reducing replication errors. XPD is involved in NER as part of the TFIIH complex making a role in NER seem feasible. We used published TFIIH cisplatin/ oxaliplatin repair sequencing (XR-seq) data [28] and found that, as with XPD ChIP-seq coverage, the mutation densities in WT samples inversely correlated with TFIIH repair, whereas the slope was largely flat for XPD mutant (-0.0712 versus 0.0067, Figure 4B) with significantly higher obs/exp and lower obs/exp in WT compared with mutant for high and low TFIIH coverage regions respectively ($q=0.000003$ and $q=0.000178$, Student's t-test with multiple testing correction, Figure S4B). The mutational spectrum in high and low TFIIH in both XPD WT and mutant cancers were also highly similar, again suggesting a loss of repair rather than XPD causing mutations (Figure S4C).

Determinants of Mutagenesis at CBS in Cancer

The previous results suggest that a loss of XPD's repair function contributes to mutation hotspots at CBS. In order to confirm statistically that the presence of XPD is a determinant of mutagenesis at CBS, as well as shed light on the potential mechanism, we performed logistic regression predicting if a CBS is mutated or not in XPD mutant cancer based on several features known to impact mutagenesis or be associated with XPD. We also performed this analysis on

other cancer types with reported mutation hotspots at CBS. All mutations from ESAD and MELA were chosen as other cancer types because samples frequently display strong SBS17 and SBS7 signatures respectively, which are associated with CBS hotspots. XPD interacts with POLR2A [29] and POLR2A ChIP-seq signal is also elevated at CBS [30, 31]. POLR2A may have an impact on local mutation densities through the recruitment of TC-NER. Genes, euchromatin and early replicating regions are typically protected from mutations. As such, we selected the following features for our model: whether or not the CBS falls within a gene, and the average replication time, XPD ChIP, CTCF ChIP, DNase hypersensitivity and POLR2A ChIP signal of the CBS +/- 150 bp.

In all cancer types, CTCF ChIP and DHS coverage significantly increased the chance of a CBS being mutated and genic CBS were less likely to be mutated (Figure 5A-C) as expected. In both ESAD and MELA replication time reduces the chance of a CBS being mutated as expected (Figure 5B-C). Replication time was not a predictor for XPD mutant cancer which is interesting as we found that XPD mutant BLCA lose replication time dependence of mutation density compared with WT (Figure 2B). XPD was not a predictor of CBS mutagenesis in XPD mutant cancer (Figure 5A). This is likely because XPD mutant cancers have lost their dependence on XPD related repair which is consistent with Figure 4B. Interestingly, XPD ChIP coverage significantly decreased the likelihood that a CBS is mutated in ESAD (OR=0.88363, pvalue =0.00218, Figure 5B). This suggests that XPD is important in the repair of SBS17 related damage. XPD was not a predictor of CBS mutagenesis in melanoma which we think points to an unrelated mechanism. We next analysed the impact of XPD coverage on ESAD mutation densities genome-wide. Since SBS17 mutations are T>G and the other predominant mutation type in ESAD is C>T, we analysed the mutation densities of these mutations with respect to

XPD ChIP and TFIIH XR-seq coverage. We found a strong inverse relationship between XPD and TFIIH XR-seq coverage and ESAD mutation densities (Figure 5D, Figure S5) for both C>T and T>G mutations. Interestingly, the slope for T>G mutations was greater than that for C>T mutations (XPD ChIP, -0.5536 vs -0.2959 and TFIIH repair, -0.2423 vs -0.1266). This suggests that XPD does indeed have a role in repair of CBS mutations.

Discussion

In this study, we present strong evidence that XPD mutant cancers display somatic mutation hotspots in CTCF-cohesin binding sites. While the biological mechanisms of this observation remain unresolved, our results suggest that it is likely related to a loss of a protective role of the WT XPD protein and implicate a role for the XPD in protecting genomic integrity of the accessible genome. While we acknowledge that it is possible that the mutant XPD protein could actively cause errors or damage in the genome, rather than the WT XPD protein preventing them under normal circumstances, we believe our results support the latter. We observed an inverse relationship between XPD coverage and mutation densities in WT BLCA samples which is lost in mutant samples. This indicates that XPD is playing a protective role in the WT samples and is absent in mutants (Figure 4B). Further, the mutational signatures were highly similar in high and low XPD regions indicating that a similar mutational process is taking place where XPD is bound compared to where XPD is not bound (Figure 4C). The fact that the mutation spectrum is unchanged, but the mutation burden increased is consistent with a loss of DNA repair. These results are also consistent with the clinical diseases progression of patients with xeroderma pigmentosum caused by XPD mutation where the incidence of skin cancer is greatly increased in

sunlight exposed skin [32]. This supports that mutant XPD does not directly generate DNA damage, but rather, is unable to repair damage arising from genomic insults such as UV-light.

Probing determinants of mutagenesis at CBS in XPD mutant, SBS7 and SBS17 related cancers further support that XPD mutant cancers lose dependence on XPD related repair and suggest that XPD may be involved in the repair of SBS17 related mutations. ESAD often display SBS17 [33], and we found that ESAD local mutation densities were inversely correlated with XPD coverage, and the slope was greater for T>G than C>T mutations, leading us to the hypothesis that XPD is involved in SBS17 related CBS hotspots. The patterns of mutagenesis across the CTCF motif of CBS in XPD mutant cancers and ESAD is similar but different to that of MELA. We believe it is unlikely that SBS17 and XPD mutant cancers could have different mechanisms leading to the exact same sites of the motif being mutated. SBS17 is believed to be a result of oxidative stress generated by gastric reflux [34], a process which causes 8-oxo-guanine misincorporation, resulting in T>G mutations [35]. Interestingly, XPD mutant cells are sensitive to oxidative stress and hence thought to have a role in the mitigation of oxidative damage [36]. We therefore postulate that SBS17 cancers are dependent on XPD related DNA repair. Our results showing decreased mutagenesis with more XPD coverage in ESAD (Figure 5B) further support this. Our finding that T>G and T>C mutations are enriched in XPD mutant CBS compared with flank (Figure 3E) is consistent with this hypothesis as these are the predominant mutations in SBS17.

It is therefore likely that accessible chromatin, particularly CBS are highly dependent on XPD related DNA repair. We speculate that this could be because transcriptionally active CBS are susceptible to strand breaks [37]. This observed dependence on DNA repair was initially surprising in light of our previous work that showed that XPC mutant UV associated skin

cancers display an absence of mutation hotspots compared with WT counterparts [17]. At that time, we concluded that CBS are vulnerable to mutation hotspots in skin cancer because they are deficient in NER. While this seems to contradict our present finding, two important points that demonstrate that these hypotheses are compatible. Firstly, XPD is required for both transcription-coupled and global NER, whereas XPC only functions in global NER, so it is possible that the damage being repaired in BLCA is dependent on TC-NER whereas in skin cancer it is global NER dependent. Logically it makes sense that transcriptionally active regions of the genome are dependent on TC-NER. However, the majority of CBS are non-coding, making the function of TC-NER in these regions less clear. The next point relates to the position of mutation hotspots within the CTCF motif. In XPD mutant and SBS17 cancers, mutations are concentrated on the left side of the motif, while in skin cancer the mutations are concentrated at the 3' GG pair in the motif. In this way, it is possible that CBS are both highly dependent on and deficient in NER on different sides of the CTCF motif. Additionally, it was recently demonstrated that UV causes more damage to the specific hotspot in the CBS motif [38], indicating a distinct mutational process from that of XPD mutant and SBS17 cancers.

Euchromatin, genic and early replicating regions of the genome display lower mutation burden than less accessible, gene rich and late replicating areas of the genome due to mismatch repair [9]. Our results suggest that XPD related DNA repair is another potential mechanism by which these regions of the genome have a lower mutation burden.

Methods

Sample cohorts

Somatic mutation calls from 23 The Cancer Genome Atlas (TCGA) whole-genome sequenced (WGS) BLCA samples were accessed from Pan-cancer analysis of whole genomes (PCAWG) [39]. Samples were defined as XPD mutant if they harboured a missense XPD mutation that is defined as 'oncogenic' or 'likely oncogenic' in OncoKB [40]. A total of 4 samples were characterised as mutant of which 2 had N238S, 1 had S44L and 1 had T484M XPD mutations (Supplementary table 1). The remaining 19 samples were classed as wild-type (WT).

To provide more evidence for observations made from the TCGA cohort, 2 Chinese cohorts of BLCA were accessed and other cancer types from PCAWG were searched for other samples with XPD mutations. In one cohort, there were 2 samples out of 65 containing recurrent, functional XPD mutations [25] and in the cohort of 6 neuroendocrine BLCA samples, 1 sample had a recurrent, functional XPD mutation [26]. In PCAWG, there were 3 liver cancer samples with recurrent, functional XPD mutations. Analysis performed on 'XPD mutant cancer' included all SNVs pooled from the above 3 cohorts of BLCA and the 3 PCAWG liver cancer samples, giving a total of 10 XPD mutant WGS cancer samples (Supplementary table 1).

Where TCGA whole exome sequenced (WXS) BLCA samples were used for additional evidence to support findings from WGS, such as in CBS contingency analysis, samples that had WGS data were excluded. WXS with XPD mutations that were 'oncogenic' or 'likely oncogenic' in OncoKB [40] were assigned as mutant as described above for WGS samples, but there were an additional 10 samples with missense mutations with a VEP score indicated pathogenic but were not previously annotated in OncoKB. One of these mutations, E606Q, was recurrent, and so we considered it a putative driver and assigned it as 'mutant'. 9 remaining samples with pathogenic,

missense mutations that were not annotated in OncoKB and not recurrent were excluded from analysis as they could not be confidently classified (Supplementary table 2). This resulted in an additional 23 and 356 XPD mutant and WT WXS samples, respectively (Supplementary table 2).

For SBS17 and SBS7 cancers we used all PCAWG esophageal adenocarcinoma (ESAD) and melanoma (MELA), respectively. This resulted in 2,728,301 mutations from 98 samples for ESAD and 12,568,609 mutations from 107 samples for MELA.

Somatic mutations and simulation

For bladder WGS TCGA samples, single nucleotide variant (SNV) calls were obtained previously [11, 41] 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). Mutation simulations were performed 100 times to calculate the distribution of mutations expected by chance based on the sequence composition of a region and the overall burden of mutations in each sample. Briefly, the trinucleotide of each position of the reference genome hg19 was defined using the R package 'BSgenome.Hsapiens.UCSC.hg19'. Then each mutation in each sample was shuffled to a random place in the genome with the same trinucleotide context using the Unix command 'shuf' 100 times.

The total number of small insertions and deletions were calculated from TCGA-BLCA.mutect2_snv.tsv.gz accessed from the Xena browser [42] as mutations where the length of the alternate and reference base was greater than 1 respectively. Structural variants were calculated from the following files downloaded from icgc portal - final_consensus_sv_bedpe_passonly.tcga.public.tgz.

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' [43]. Mutations of 100 simulations were merged for analysis and then divided by 100 to give an 'expected' value, and then local mutation density was expressed as the ratio of observed to expected mutations. These analyses require a minimum mutation burden in each sample in order to be able to generate reliable ratios. Many of the WT samples had low mutation numbers, particularly for Other mutations, and some samples had low mutation numbers for APOBEC processes. Rather than completely discarding lowly mutated samples, we pooled the mutations of these samples together and displayed them graphically in analysis as indicated in figure legends. The lowest Other mutation count in the XPD mutant samples was 8219, is sufficient for analysis [11], and we therefore used a threshold of 8000 as the minimum number of Other and APOBEC mutations a sample must have to be included with samples lower than this being pooled. This pooled sample was not included in t-tests but was displayed graphically and almost always fell on the median, illustrating that these lowly mutated samples behave in the same way as the others in the cohort. IDs of these lowly mutated samples used for pooling are provided in supplementary table 3.

Genome wide distribution of mutations was performed by calculating mutation densities as described above for 1 megabase (mb) windows of hg19 and then principal component analysis (PCA) was performed 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 each site of the region separately 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

To calculate trinucleotide frequencies, 'slopBed' [43] was used to extend 1 base on either side of the SNV followed by 'fastafrombed' [43] for hg19 to retrieve the trinucleotide context of mutations. The total of each trinucleotide mutation was counted and divided by the total number of mutations to obtain its frequency. For the specific genomic regions of which the mutation trinucleotide frequencies were to be calculated, these values were divided by the trinucleotide counts of the region then multiplied by the trinucleotide counts for the whole genome in order to perform a region to genome normalisation.

Genomic Annotations and Data Binning

Gene expression data was taken from 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 [44] (ENCSR813CKU) (Supplementary table 4) as bigWig and bed file respectively. ChIP-seq for

XPD, XPB and input was accessed from GEO (GSE44849) and cisplatin/ oxoplatin based TFIIH XR-seq was accessed from GEO under accession GSE82213 as bigwig files. Deeptools 'bigWigCompare' was used to generate ChIP to input log2 ratio bedgraph files for XPD and XPB ChIP-seq. For TFIIH XR-seq data, 'bigWigMerge' was used to merge plus and minus outputting a bedgraph. 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, XPD ChIP-seq and TFIIH, the mean bedgraph signal from genomic assays was calculated for each of the 1kb filtered genomic windows using bedtools map [43]. 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) (Supplementary table 4) and then gene start sites to get promoters. Bladder DHS peaks were overlapped with high quality, experimentally determined CBS accessed from supplementary materials of [14] to generate CBS annotations. Later analysis of CBS uses these high quality CBS annotations [14] without overlapping with DHS. Finally, enhancers were defined as the centre of bladder H3K27Ac ChIP-seq peaks from ENCODE (ENCSR054BKO) (Supplementary table 4) that overlapped bladder DHS peaks. Chromatin A/B compartments for bladder were taken from supplementary files of [45]. For later analysis on CBS, all 31252 CBS sites were used [14].

Generating Coverage Profiles Across Genomic Regions for ChIP-seq Data

BigWig files for XPD and XPB ChIP-seq and input were accessed from gene expression omnibus (GEO) under accession GSE44849 which was previously published [27] (Supplementary table 4). Deeptools 'bigwigCompare' [46] was used to generate log2 ratio bigwig files of the ChIP compared with input, skipping regions that had no coverage in both the input and the ChIP. To generate sequencing coverage profile plots for regions, windows were generated within, upstream or downstream of each site in the region 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. UCSC tool 'bigWigAverageOverBed' was then used to retrieve the average signal of each region in each window. The average region signal of each window was averaged and plotted positionally.

Regression Analyses

Logistic regressions were performed to predict the chance of a site being mutated. We modelled whether a CBS is mutated or not based on the replication time, CTCF ChIP, XPD ChIP, DHS and POLR2A ChIP signal of that CBS and whether it fell in a gene body in XPD mutant cancer, ESAD and MELA.

Whether the site was mutated in the given cancer was calculated using intersectBed [43]. Then the average signal for each of these assays was calculated using bedtools map [43] on bedgraph files, mapping to each of the sites. As CBS regions were only 39 bp, we extended the region by 150 bp on each side using slopBed [43] before mapping bedgraph signal.

In order to generate the aforementioned bedgraph files consistently, we downloaded filtered hg19 bam files from ENCODE for ChIP and DHS data (Supplementary table 4). For ENCODE

ChIP-seq experiments, deeptools bamCompare [46] was used to generate bedgraph files of ChIP normalised to respective input files as log2 ratios. XPD ChIP-seq was accessed from GSE44849 (Supplementary table 4) and deeptools bigWigCompare was used to generate ChIP signal as L2R to input. For DHS, deeptools bamCoverage [46] was used skipping mitochondria and output RPKM signal. Replication time was obtained from the UCSC table browser (wgEncodeEH002244). All of the above variables were z-score normalised. As an additional binary variable, whether sites were genic or intergenic was determined using intersectBed.

Out of 31252 CBS sites, a total of 1283, 2959 and 5902 sites were mutated for XPD mutant cancer, ESAD and MELA respectively. In order to balance the mutated and non-mutated sites, 5%, 20% and 10% of non-mutated sites were randomly extracted 100 times for XPD mutant cancer, esophageal adenocarcinoma and melanoma respectively. Logistic regression was performed in R. Adjusted odds ratios and p values were calculated based on a multivariable model.

Contributions: JAB analysed data, prepared figures, and wrote the manuscript; TO, HF, NCY, XZ, HY, YTW analysed and interpreted data, MWW, NAB contributed data, SW contributed data analysis, study design, and supervised research, J.W.H.W. conceived and designed the study, analysed data, revised manuscript and supervised the research. All authors read and approved the final manuscript.

Acknowledgements: The project was supported by the Research Grants Council, HK (17100920, R7022-20 and C7028-19G) to JWHW.

Competing interests: The authors have declared that no competing interests exist.

Data and materials availability: Links to dataset used in the paper are listed in the Supplementary Tables. Source code to all scripts related to the data analysis are available on GitHub at https://github.com/jayneAbarbour/final_ERCC2_paper

References

1. Compe, E. and J.M. Egly, *TFIIH: when transcription met DNA repair.* Nat Rev Mol Cell Biol, 2012. **13**(6): p. 343-54.
2. Singh, A., et al., *TFIIH subunit alterations causing xeroderma pigmentosum and trichothiodystrophy specifically disturb several steps during transcription.* Am J Hum Genet, 2015. **96**(2): p. 194-207.
3. Van Allen, E.M., et al., *Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma.* Cancer Discov, 2014. **4**(10): p. 1140-53.
4. *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-22.
5. Li, Q., et al., *ERCC2 Helicase Domain Mutations Confer Nucleotide Excision Repair Deficiency and Drive Cisplatin Sensitivity in Muscle-Invasive Bladder Cancer.* Clin Cancer Res, 2019. **25**(3): p. 977-988.
6. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. **578**(7793): p. 94-101.
7. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
8. Schuster-Böckler, B. and B. Lehner, *Chromatin organization is a major influence on regional mutation rates in human cancer cells.* Nature, 2012. **488**(7412): p. 504-7.
9. Supek, F. and B. Lehner, *Differential DNA mismatch repair underlies mutation rate variation across the human genome.* Nature, 2015. **521**(7550): p. 81-4.
10. Frigola, J., et al., *Reduced mutation rate in exons due to differential mismatch repair.* Nat Genet, 2017. **49**(12): p. 1684-1692.
11. Perera, D., et al., *Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes.* Nature, 2016. **532**(7598): p. 259-63.
12. Sabarinathan, R., et al., *Nucleotide excision repair is impaired by binding of transcription factors to DNA.* Nature, 2016. **532**(7598): p. 264-7.
13. Poulos, R.C., J. Olivier, and J.W.H. Wong, *The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes.* Nucleic Acids Res, 2017. **45**(13): p. 7786-7795.
14. Katainen, R., et al., *CTCF/cohesin-binding sites are frequently mutated in cancer.* Nat Genet, 2015. **47**(7): p. 818-21.
15. Hnisz, D., et al., *Activation of proto-oncogenes by disruption of chromosome neighborhoods.* Science, 2016. **351**(6280): p. 1454-1458.
16. Kaiser, V.B., M.S. Taylor, and C.A. Semple, *Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types.* PLoS Genet, 2016. **12**(8): p. e1006207.
17. Poulos, R.C., et al., *Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif.* Cell Rep, 2016. **17**(11): p. 2865-2872.
18. Guo, Y.A., et al., *Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers.* Nat Commun, 2018. **9**(1): p. 1520.
19. Kaiser, V.B. and C.A. Semple, *Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline.* Genome Biol, 2018. **19**(1): p. 101.

20. Holwerda, S.J. and W. de Laat, *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120369.
21. Kim, J., et al., *Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors.* Nat Genet, 2016. **48**(6): p. 600-606.
22. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nat Genet, 2013. **45**(9): p. 970-6.
23. Seplyarskiy, V.B., et al., *APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication.* Genome Res, 2016. **26**(2): p. 174-82.
24. Lada, A.G., et al., *Disruption of Transcriptional Coactivator Sub1 Leads to Genome-Wide Re-distribution of Clustered Mutations Induced by APOBEC in Active Yeast Genes.* PLoS Genet, 2015. **11**(5): p. e1005217.
25. Wu, S., et al., *Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications as angiogenesis-related drivers in bladder cancer.* Nat Commun, 2019. **10**(1): p. 720.
26. Shen, P., et al., *Comprehensive genomic profiling of neuroendocrine bladder cancer pinpoints molecular origin and potential therapeutics.* Oncogene, 2018. **37**(22): p. 3039-3044.
27. Gray, L.T., et al., *G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD.* Nat Chem Biol, 2014. **10**(4): p. 313-8.
28. Hu, J., et al., *Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution.* Proc Natl Acad Sci U S A, 2016. **113**(41): p. 11507-11512.
29. Peissert, S., et al., *In TFIIH the Arch domain of XPD is mechanistically essential for transcription and DNA repair.* Nat Commun, 2020. **11**(1): p. 1667.
30. Ramanand, S.G., et al., *The landscape of RNA polymerase II-associated chromatin interactions in prostate cancer.* J Clin Invest, 2020. **130**(8): p. 3987-4005.
31. Zhang, S., et al., *RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis.* Sci Adv, 2021. **7**(43): p. eabg8205.
32. Rizza, E.R.H., et al., *Xeroderma Pigmentosum: A Model for Human Premature Aging.* J Invest Dermatol, 2021. **141**(4s): p. 976-984.
33. Secrier, M., et al., *Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance.* Nat Genet, 2016. **48**(10): p. 1131-41.
34. Dvorak, K., et al., *Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus.* Gut, 2007. **56**(6): p. 763-71.
35. Suzuki, T. and H. Kamiya, *Mutations induced by 8-hydroxyguanine (8-oxo-7,8-dihydroguanine), a representative oxidized base, in mammalian cells.* Genes Environ, 2017. **39**: p. 2.
36. Lerner, L.K., et al., *XPD/ERCC2 mutations interfere in cellular responses to oxidative stress.* Mutagenesis, 2019. **34**(4): p. 341-354.
37. Canela, A., et al., *Topoisomerase II-Induced Chromosome Breakage and Translocation Is Determined by Chromosome Architecture and Transcriptional Activity.* Mol Cell, 2019. **75**(2): p. 252-266.e8.

38. Sivapragasam, S., et al., *CTCF binding modulates UV damage formation to promote mutation hot spots in melanoma.* Embo j, 2021. **40**(20): p. e107795.
39. *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
40. Chakravarty, D., et al., *OncoKB: A Precision Oncology Knowledge Base.* JCO Precis Oncol, 2017. **2017**.
41. Poulos, R.C., et al., *Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations.* PLoS Genet, 2018. **14**(11): p. e1007779.
42. Goldman, M.J., et al., *Visualizing and interpreting cancer genomics data via the Xena platform.* Nat Biotechnol, 2020. **38**(6): p. 675-678.
43. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
44. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-d801.
45. Fortin, J.P. and K.D. Hansen, *Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data.* Genome Biol, 2015. **16**(1): p. 180.
46. Ramírez, F., et al., *deepTools: a flexible platform for exploring deep-sequencing data.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W187-91.

Figures

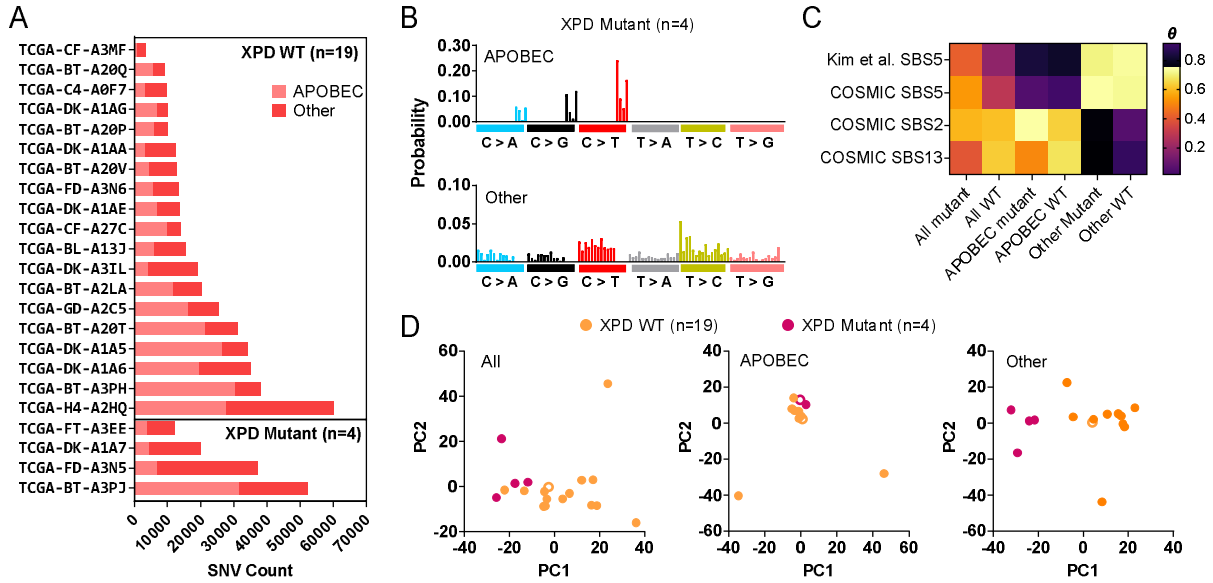

Figure 1 Contribution of APOBEC and Other Mutations in XPD mutant and WT Bladder Cancer (A) Total number of mutations attributed to T[C>D]N (APOBEC) or not T[C>D]N (Other) in each sample of the TCGA WGS bladder cancer cohort arranged by genotype and total mutation number. (B) Trinucleotide mutational spectra of APOBEC and Other mutations for TCGA bladder cancer samples. (C) Heat map of cosine similarities to mutational signatures in bladder cancer. COSMIC signatures 2, 5 and 13 are from COSMIC, the other signature, TCGA.130.DFCI.MSK.50.signature5 is from the supplementary material from [21] (D) Principle component (PC) analysis plots representing PC1 and PC2 of observed-expected mutation density ratios were calculated across each 1 mb window of hg19 for all SNVs, APOBEC SNVs and Other SNVs

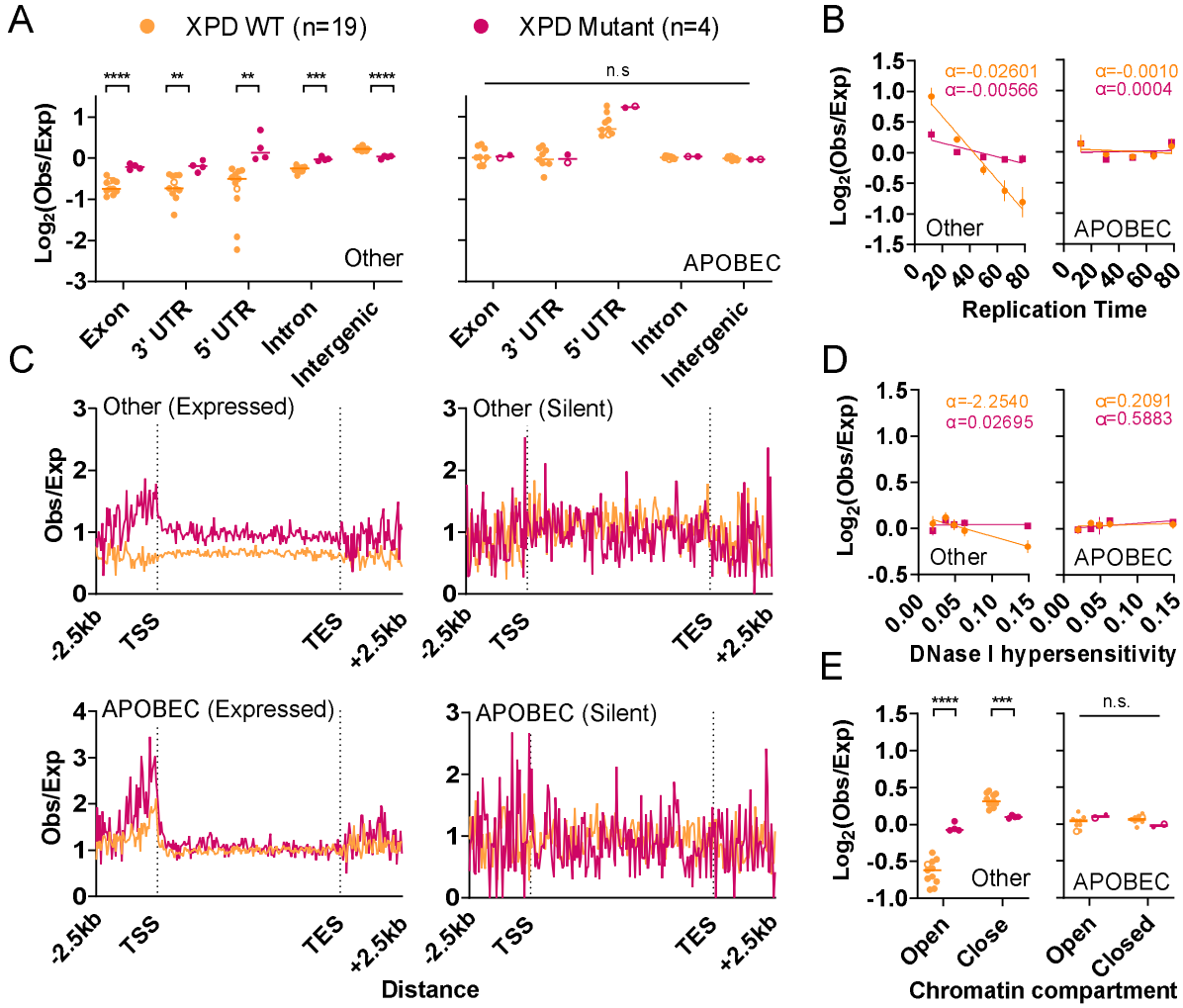

Figure 2 Genome-Wide distribution of APOBEC and Other Mutations in XPD mutant and WT Bladder Cancer (A) Mutation densities as observed-expected ratios (obs/exp) in exons, and 3' and 5' untranslated regions (UTR), introns or not in any of these regions (intergenic) in WT samples in orange and XPD Mutant samples (Mutant) in pink with Other SNVs displayed on the left and APOBEC SNVs displayed on the right. Large hollow circle point represents the pooled mutations of lowly mutated samples. ** q < 0.01, *** q < 0.001, **** q < 0.0001, n.s. not significant, Student's t-test with multiple testing correction. (B) Mutation densities as obs/exp for 5 genomic bins organised by 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. (C) Observed-expected mutation density ratio profile plots Other SNVs (left) and APOBEC SNVs (right) across gene body 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 organised into 150 bins and the region 2.5 Kb up or downstream of the TSS or TES was organised into 50 bins. (D) Mutation densities as obs/exp for 5 genomic bins organised by DNase hypersensitivity (DHS). 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. (E) Observed/expected mutation density ratios for genomic regions annotated in normal bladder to be either a chromatin A compartment (open) or chromatin B compartment (closed). Large hollow circle point represents the pooled mutations of lowly mutated samples

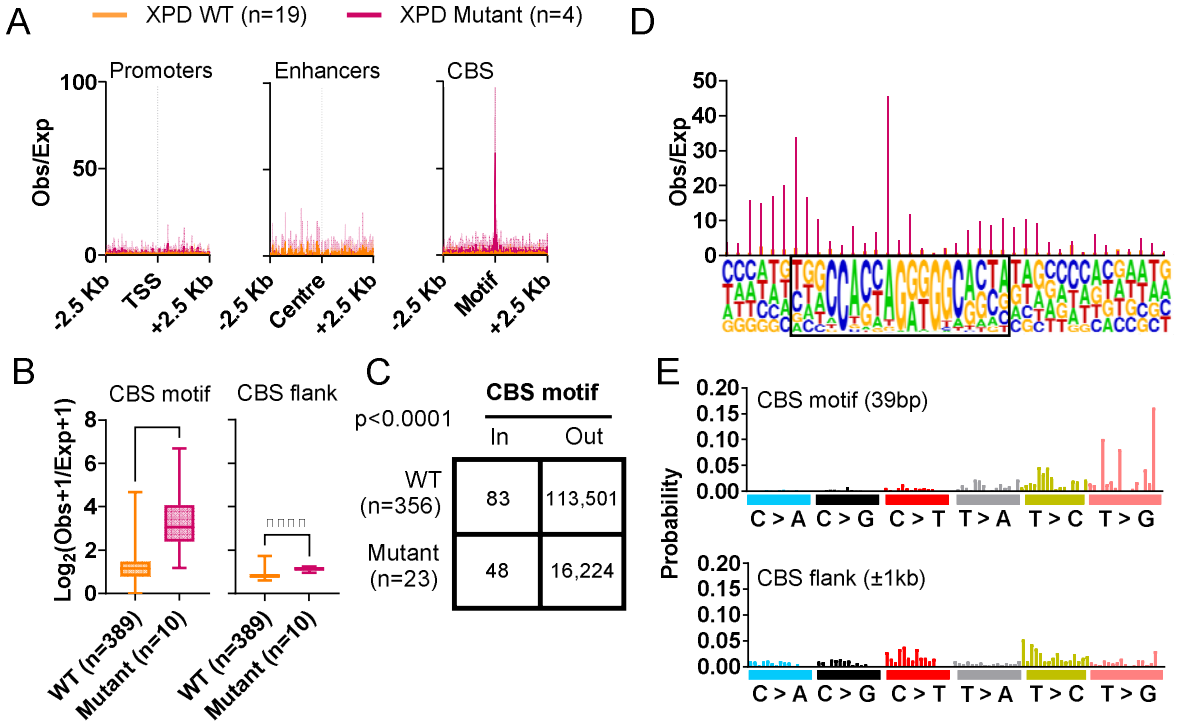

Figure 3 Mutation Densities at DNase Hypersensitive Regions in XPD mutant and WT Bladder Cancer. (A) Profile plots of mutation densities as observed-expected ratios (obs/exp) for Other Mutations in regions annotated in normal bladder as the promoter, enhancer or CTCF-Cohesin binding site (CBS) 2.5 Kb up or downstream of the TSS, centre and motif, respectively. (B) Mutation densities (obs/exp) of CBS motif and +/-1 Kb flanking regions for all XPD mutant cancer samples (n=10). **** q < 0.0001, Student's t-test with multiple testing correction. (C) Contingency table displaying the number of mutations falling in CTCF-Cohesin binding sites (CBS) or any other region of the exome (Other) in XPD mutant and WT TCGA whole-exome sequenced bladder cancer samples that were not also whole-genome sequenced. (D) Observed/expected mutational profile of TCGA XPD mutant samples across the CBS motif. (E) Trinucleotide mutation frequencies of bladder XPD mutations in CBS and flanking regions.

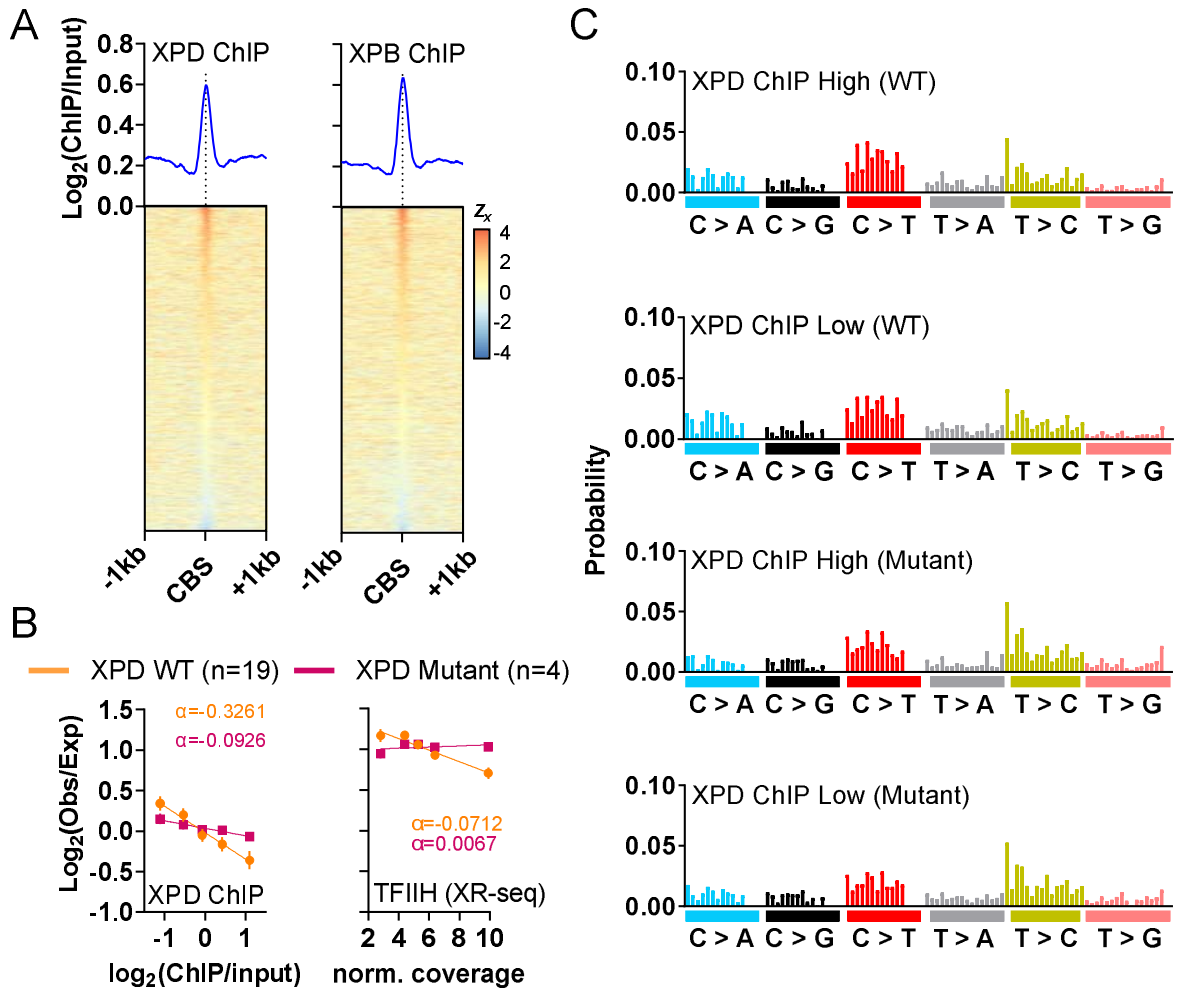

Figure 4 Effect of XPD on Mutagenesis in WT and XPD Mutant Bladder Cancer

(A) Profile and heat maps of coverage of XPD and XPB ChIP-seq across CBS. Data accessed from (GSE44849). (B) Mutation densities in genomic bins with respect to Mutation densities as obs/exp for 5 genomic bins organised by XPD ChIP and TFIIH cisplatin/oxaliplatin XR-seq coverage. Plots and error bars represent the mean and standard deviation of different samples. The line represents a linear regression model between mutation densities and the mean coverage for each of the bins. (C) Trinucleotide mutation frequencies (fraction of total mutations) for all mutations falling in high and low XPD ChIP coverage regions for WT and XPD mutant bladder cancer. These frequencies are scaled by the average trinucleotide composition of the regions.

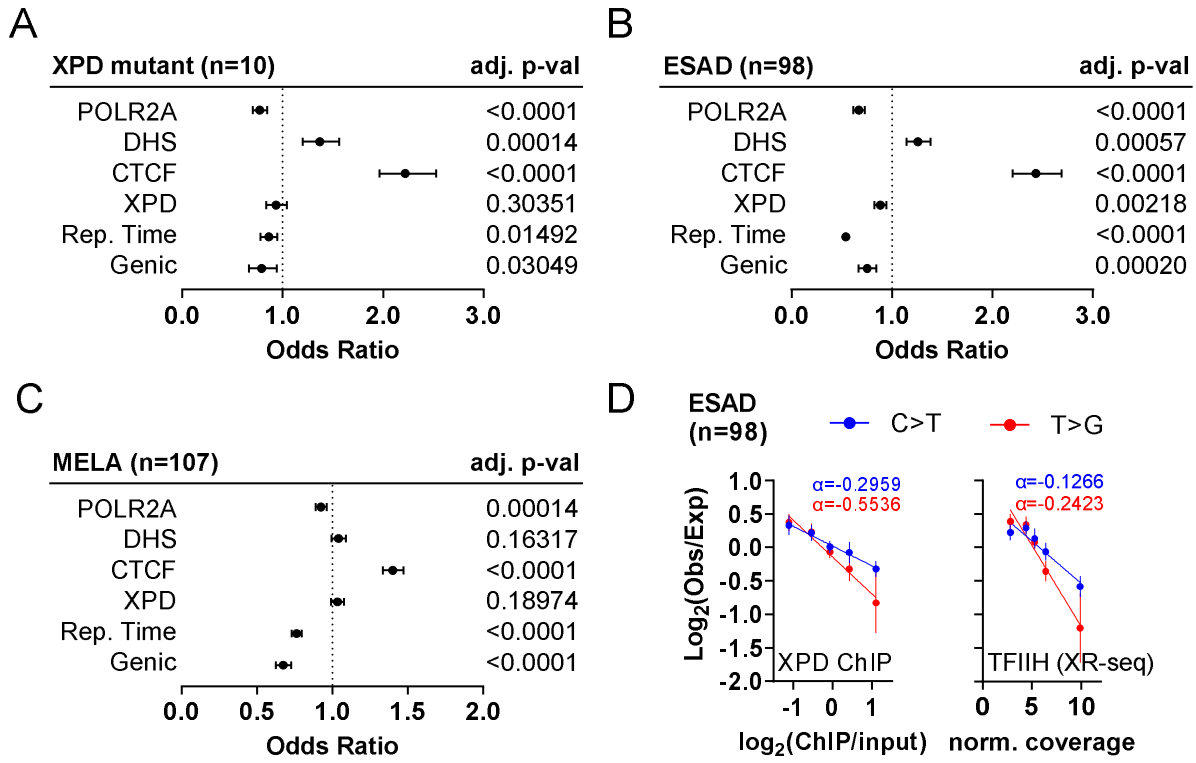

Figure 5 Determinants of Mutagenesis at CTCF-Cohesin Binding Sites. Multivariable logistic regression models predicting whether a CBS is mutated or not based on genic and epigenetic features for (A) XPD mutant cancers (XPD Mutant), (B) SBS17 esophageal adenocarcinoma (ESAD) and (C) SBS7 related melanoma (MELA). (D) PCAWG esophageal adenocarcinoma C>T (blue) and T>G (red) mutation densities in genomic bins with respect to observed/expected mutations for 5 genomic bins organised by XPD ChIP and TFIIH cisplatin/oxaliplatin XR-seq coverage. Plots and error bars represent the mean and standard deviation of different samples. The line represents a linear regression model between mutation densities and the mean coverage for each of the bins.

Supplementary Figures

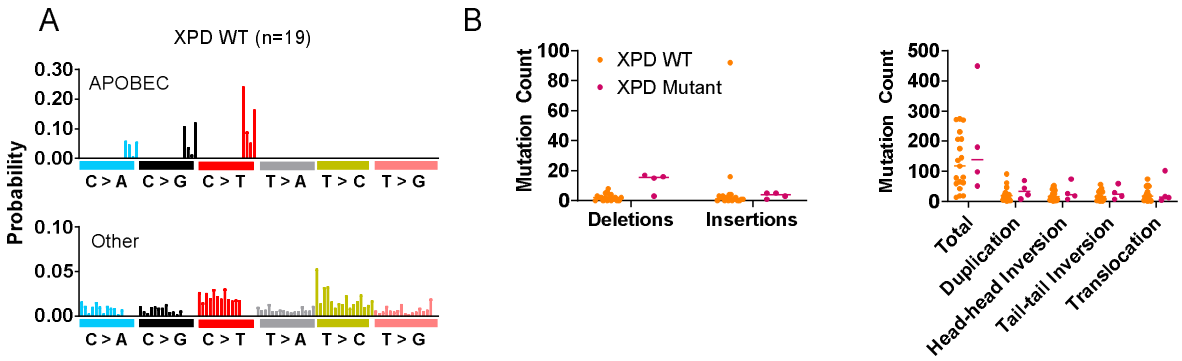

Figure S1 – Contribution of different mutations in XPD mutant and wild-type (WT) bladder cancer. (A) Weighted trinucleotide mutation frequencies (fraction of total mutations) for mutations attributed to T[C>D]N trinucleotide (APOBEC), and mutations not attributed to T[C>D]N (Other) in BLCA XPD WT samples. (B) Total number of small insertions and deletions (left) and structural variants (right) per sample.

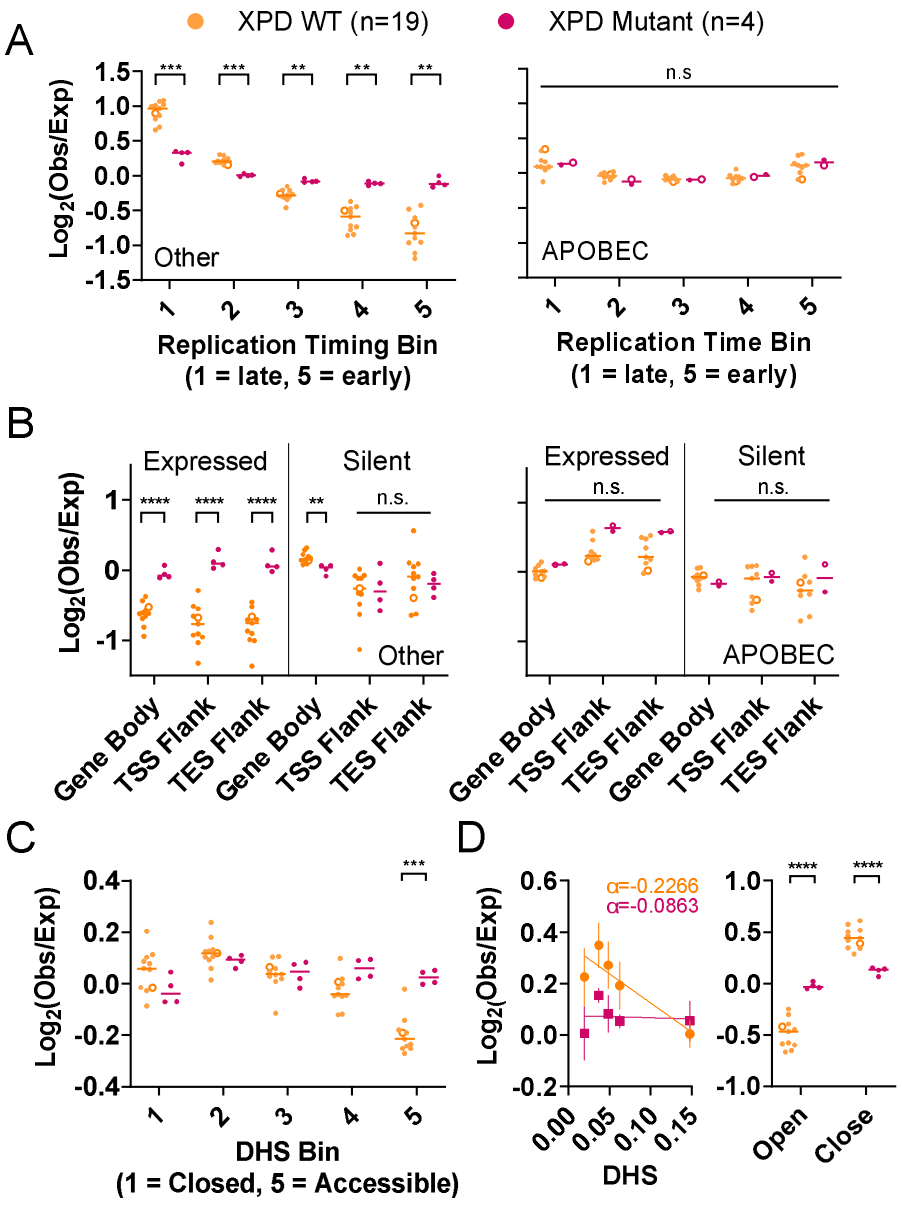

Figure S2 – Genome wide distribution of mutations in XPD Mutant and WT bladder cancer (A) Mutation densities of individual data points with respect to replication time bins for Other and APOBEC SNVs. (B) A statistical representation of plots in Fig 2C. Mutation densities of individual samples are shown for the gene body, the 2.5 Kb upstream flanking region of TSS (TSS flank) and 2.5 Kb downstream flanking region of TES (TES flank). (C) Mutation densities of individual data points with respect to DNase hypersensitivity (DHS) bins for Other and

APOBEC SNVs. (D) Analysis in figure 2D and 2E performed with CBS and genes subtracted from DHS bins and from open and closed A/B compartments. Large hollow circle point represents the pooled mutations of lowly mutated samples. ** $q < 0.01$, *** $q < 0.001$, **** $q < 0.0001$, n.s. not significant, Student's t-test with multiple testing correction.

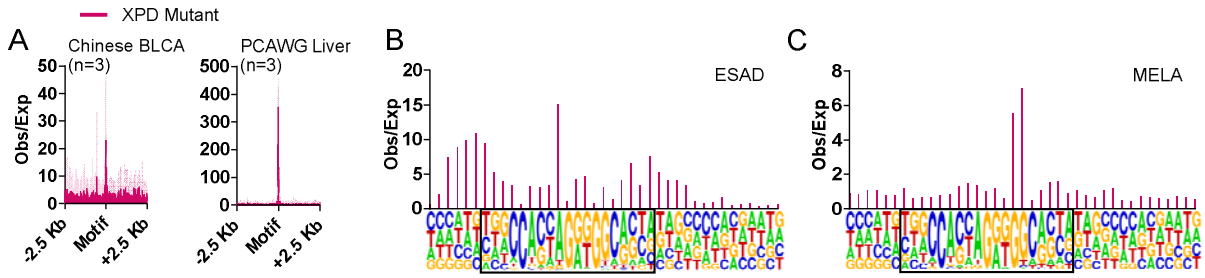

Figure S3 Mutation profiles across CBS and flanking regions. (A) Mutation profiles across CBS in XPD mutants from Chinese BLCA cohort (left) and PCAWG liver cancer samples (right). Observed/Expected mutation profile across the CTCF motif for esophageal adenomcarcinoma (ESAD) (B) and melanoma (MELA) (C).

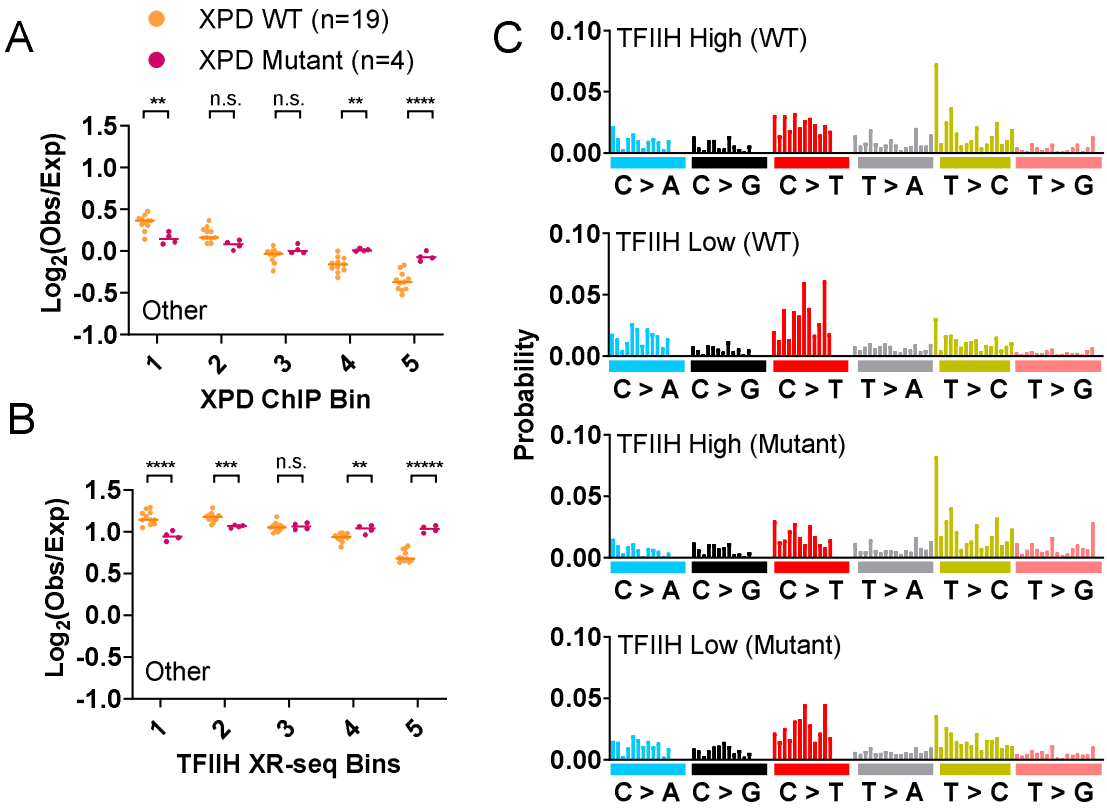

Figure S4 Mutation load and mutational signature in relation to XPD binding and TFIIH XR-seq coverage. Observed/Expected mutation load across XPD ChIP-seq (A) and TFIIH XR-seq (B) coverage bins for Other SNVs. (C) Trinucleotide mutation frequencies (fraction of total mutations) for mutations falling in high and low TFIIH XR-seq coverage regions for WT and XPD mutant bladder cancer. These frequencies are scaled by the average trinucleotide composition of the regions. ** q < 0.01, *** q < 0.001, **** q < 0.0001, n.s. not significant, Student's t-test with multiple testing correction.

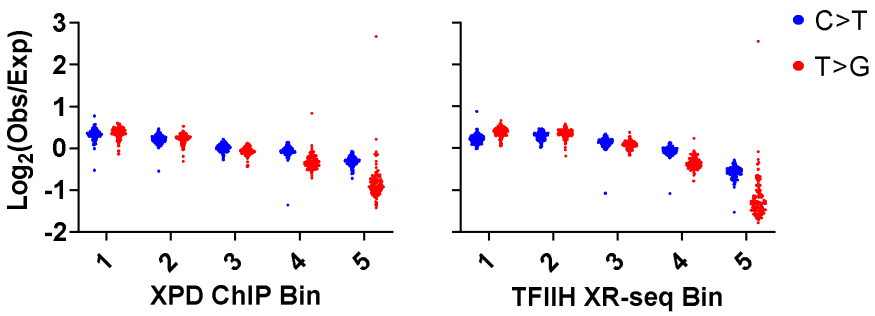

Figure S5 Effect of XPD and TFIIH repair on mutagenesis in esophageal adenocarcinoma. Observed/Expected mutation load across XPD ChIP-seq (A) and TFIIH XR-seq (B) coverage bins for C>T (blue) and T>G (red) mutations.

Supplementary Tables

Supplementary table 1. XPD mutant whole genome sequenced samples analysed in the study

Supplementary table 2. XPD mutation status of all TCGA BLCA samples

Supplementary table 3. Samples with low mutation counts that were pooled in analyses

Supplementary table 4. List of public datasets used in the study