Whole-genome mapping of APOBEC mutagenesis in metastatic urothelial carcinoma identifies driver hotspot mutations and a novel mutational signature

J. Alberto Nakauma-González[1,2,3,*], Maud Rijnders[3], Minouk T. W. Notestordsij[1], John W. M. Martens[3], Astrid A.M. van der Veldt[3,4], Martijn P. J. Lolkema[3,6], Joost L. Boormans[2], Harmen J. G. van de Werken[1,2,5,7,*]

[1]Cancer Computational Biology Center, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[2]Department of Urology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[3]Department of Medical Oncology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[4]Department of Radiology & Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[5]Department of Immunology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[6]Currently employed by Amgen Europe BV, Breda 4817 ZK, the Netherlands
[7]Lead contact

*Correspondence: j.nakaumagonzalez@erasmusmc.nl (JAN), h.vandewerken@erasmusmc.nl (HJGvdW), Cancer Computational Biology Center & Department of Immunology, Department of Urology, Erasmus MC Cancer Institute, University Medical Center; Rotterdam, the Netherlands. PO Box 2040, 3000 CA, NA-1218.

SUMMARY

APOBEC enzymes mutate specific DNA sequences and hairpin-loop structures, challenging the distinction between passenger and driver hotspot mutations. Here, we characterized 115 whole-genomes of metastatic urothelial carcinoma (mUC) to identify APOBEC mutagenic hotspot drivers. APOBEC-associated mutations were detected in 92% of mUC and were equally distributed across the genome, while APOBEC hotpot mutations (ApoHM) were enriched in open chromatin. Hairpin-loops were frequent targets of didymi (twins in Greek), two hotpot mutations characterized by the APOBEC mutational signature SBS2, in conjunction with an uncharacterized mutational context (Ap[C>T]), which was associated with DNA mismatch. Next, we developed a statistical framework that identified 0.40% of ApoHM as drivers of mUC, which affected known driver genes and non-coding regions near exons of potential novel driver genes. Our results and statistical framework were validated in independent cohorts of 23 non-metastatic UC and 3744 samples of 17 metastatic cancers, identifying cancer-type-specific drivers. Our study highlights the role of APOBEC in cancer development and may contribute to developing novel targeted therapy options for APOBEC-driven mUC.

KEYWORDS

APOBEC; breast cancer; didymi; driver mutations; hairpin-loops; hotspot mutations; pan-cancer; mutational signature; twin mutations; urothelial carcinoma

INTRODUCTION

Cancer genomes accumulate somatic mutations via different mutagenic processes and one of the most common is attributed to the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) family[1]. APOBEC has a specific mutational signature, which is characterized by C>T/G mutations in the TpC context and is captured in the SBS2 and SBS13 signatures, as defined by the *Catalogue Of Somatic Mutations In Cancer* (COSMIC)[2]. In some tumor types with high APOBEC activity, the contribution to the tumor mutational burden is substantial, which increases the neoantigen load favoring response to immune checkpoint inhibitors[3,4]. However, APOBEC is also responsible for the emergence of driver mutations that contribute to cancer development as shown in mouse models[5]. Discriminating driver events from passenger events is essential to reconstruct the evolutionary history of cancers and identify effective novel drug targets in APOBEC-driven tumors.

The mutational process of APOBEC has been extensively studied, revealing its preference for single-stranded DNA structures that form hairpin-loops[6]. This characteristic of APOBEC could result in identical somatic mutations in tumors from multiple patients, so-called hotspot mutations or hotspots. Due to their high prevalence, these hotspots can erroneously be assigned as driver mutations, especially in the non-coding area of the genome. However, the vast majority of mutations are passengers and do not contribute to cancer development, and the same principle may also apply to hotspot mutations[7]. Although bioinformatic strategies to identify driver hotspot mutations have been developed[8,9], the unique characteristics of the APOBEC mutagenic process require specific considerations to account for all co-variables accurately.

APOBEC-derived mutations are a dominant contributor to the mutational landscape in urothelial carcinoma (UC). Therefore, we analyzed whole-genome DNA-sequencing data of 115 metastatic UC (mUC) and matched blood samples[10] to identify driver hotspot mutations in the context of APOBEC mutagenesis. The comprehensive characterization of APOBEC-enriched tumors identified a novel mutational signature associated with DNA mismatch repair as well as genomic co-variates associated with APOBEC-derived hotspot mutations (ApoHM), which we used to develop a statistical framework and identify driver ApoHM. Furthermore, our findings were validated in whole-genomes of an independent cohort of 23 non-metastatic UC, and the analysis was extended to include 442 metastatic breast cancer (mBC) and 3302 samples 16 other metastatic cancer types.

RESULTS

APOBEC mutagenesis dominates the mutational landscape of urothelial carcinoma

The analysis of whole-genome sequencing (WGS) data of mUC and matched blood samples revealed a median of 20,667 (Q1=14,304, Q3=31,411) single nucleotide variants (SNVs) per tumor. mUC with a significant enrichment (E) for C>T mutations in TCW (W = A or T) context were considered APOBEC positive (92%). These tumors were further stratified according to APOBEC enrichment as APOBEC-high (41%; E>3), APOBEC-medium (33%; 2<E<3), and APOBEC-low (18%; E>1) (**Figure 1A**). The median contribution of APOBEC COSMIC signatures (SBS2+SBS13) in APOBEC-high, -medium and -low tumors was 61%, 37% and 15%, respectively. For the remaining 8% of tumors lacking APOBEC mutations, the median APOBEC signature was <2%, potentially reflecting the noise of the mutational signature calling. We associated the APOBEC stratification with multiple factors to better assess the different APOBEC subtypes. Tumor purity, for instance, declined with increasing APOBEC mutagenesis (**Figure S1**). Moreover, age was associated with the enrichment of APOBEC mutations (**Figure S1**). The median clonal fraction of SNVs was lower in tumors with APOBEC mutations than in non-APOBEC tumors, suggesting higher tumor heterogeneity in APOBEC-enriched tumors (**Figure S1**). Localized hypermutation events (kataegis) strongly correlated with APOBEC enrichment (spearman r=0.80, p<0.001). Homologous recombination (HR) deficiency (n=3) was only present in APOBEC-low tumors (Fisher's exact p=0.005), while none of the patients with microsatellite instability (MSI; n=4) had evidence of APOBEC mutagenesis (Fisher's exact p<0.001). Structural variants were more frequent in APOBEC tumors than in non-APOBEC tumors (**Figure S1**). Additionally, APOBEC tumors had a higher ploidy (median ploidy = 3) and a higher number of genes affected by copy number alterations (CNA) than non-APOBEC tumors (**Figure S1**), suggesting genomic instability in APOBEC-driven mUC tumors. APOBEC mutagenesis was not associated with sex, the primary origin of mUC (upper tract versus bladder) or chromothripsis.

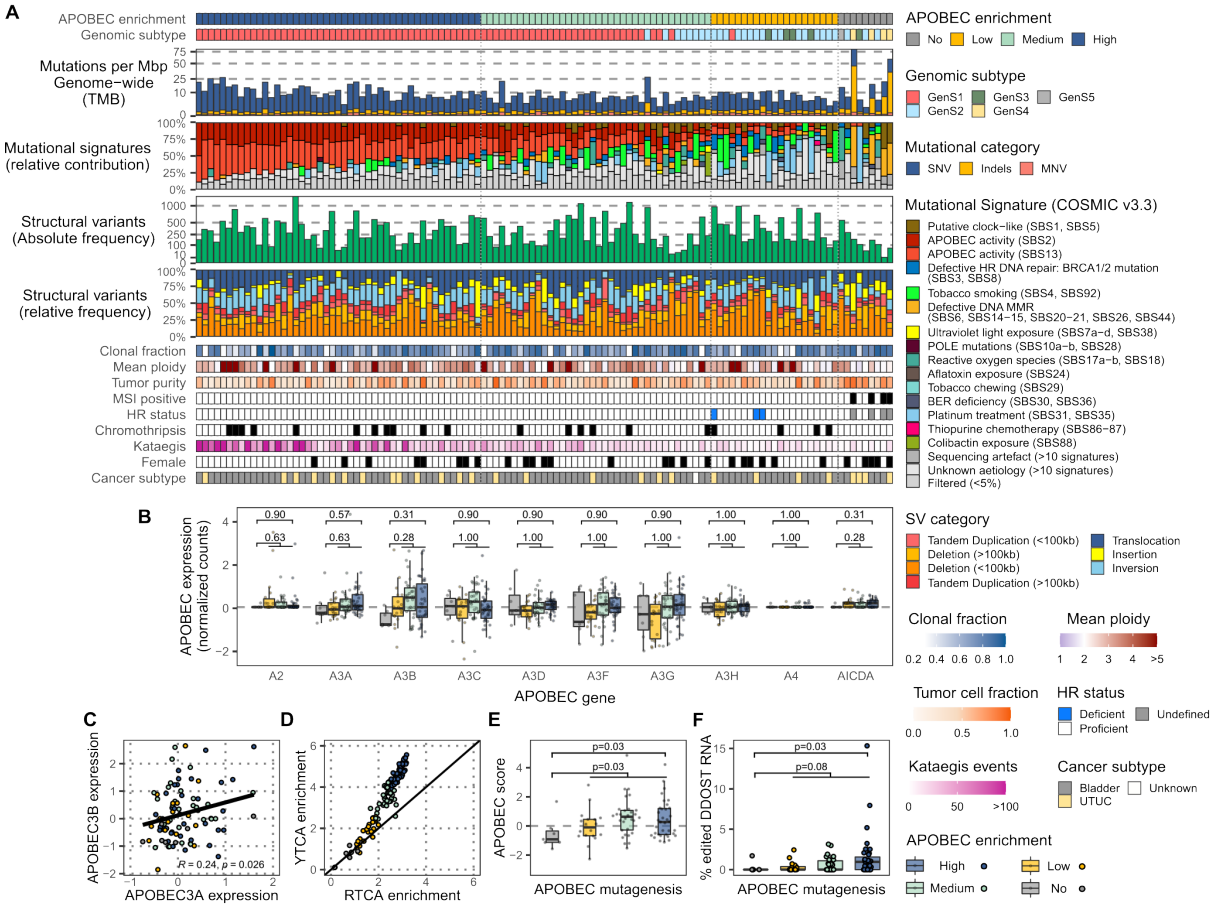


Figure 1. Genomic landscape and APOBEC activity of metastatic urothelial carcinoma (n=115) stratified by APOBEC-enrichment

(A) Whole-genome sequencing data of metastatic urothelial carcinoma (mUC) were classified according to the enrichment of APOBEC-associated mutations as having high, medium, low or no APOBEC enrichment. The genomic features are displayed from top to bottom as follows: APOBEC mutagenesis; genomic subtype (GenS1-5) as previously described[10]; genome-wide tumor mutational burden (TMB); mutational signatures grouped by etiology, while both APOBEC signatures are shown separately; absolute frequency of structural variants (SV); relative frequency of SV categories; clonal fraction; ploidy; tumor purity; microsatellite instability (MSI) status; homologous recombination (HR) deficiency status; samples with at least one chromothripsis event; frequency of kataegis events; female patients and; primary origin of mUC (upper tract versus bladder).

(B) Expression of *APOBEC* and *AICDA* genes in 90 samples with available RNA-sequencing data.

(C) Pearson correlation of RNA expression of *APOBEC3A* and *APOBEC3B.*

(D) Fold enrichment of C>T and C>G alterations in YTCA (related to APOBEC3A) and RTCA (related to APOBEC3B) context.

(E) APOBEC score (normalized expression of *APOBEC3A*+*APOBEC3B*).

(F) Percentage of mRNA C>U mutations in *DDOST* at position chr1:20981977. In (B), (E) and (F), the Wilcoxon rank-sum test was applied to compare APOBEC tumors *vs.* non-APOBEC tumors. P-values were Benjamini-Hochberg corrected in (B).

See also Figures S1-S5.

APOBEC mutagenesis is an ongoing process in metastatic lesions of urothelial carcinoma

Next, we analyzed RNA-sequencing data of 90 matched samples of mUC. Pathway activity based on downstream gene expression, such as cell cycle or p53 was similar between the APOBEC groups (**Figure S2**). Similarly, analysis of APOBEC expression of all genes of the APOBEC family (*APOBEC1* was not expressed) revealed no significant differences between APOBEC and non-APOBEC tumors (**Figure 1B**). We detected a weak positive correlation between the expression of *APOBEC3A* and *APOBEC3B* (**Figure 1C**). To further investigate the mutagenic activity of both enzymes, the fold enrichment of C>T and C>G mutations, at DNA level, in the tetra-base YTCA (related to APOBEC3A; Y are pyrimidine bases) and RTCA (related to APOBEC3B; R are purine bases) context was calculated[11] (**Figure 1D**). In both cases, YTCA and RTCA mutations did not correlate with expression of *APOBEC3A* or *APOBEC3B* (**Figure S3**). The lack of correlation might be linked to the heterogeneous expression of APOBEC enzymes that oscillate throughout the cell cycle[12,13]. Furthermore, we detected that both, APOBEC3A and APOBEC3B, contributed to APOBEC-associated mutations (fold enrichment is above 1.0). Nevertheless, APOBEC3A appeared to be the main contributor as suggested in primary cancers[11,14]. Considering the mRNA expression of both *APOBEC3A* and *APOBEC3B* enzymes, we calculated the APOBEC expression score (sum of the normalized expressions of *APOBEC3A* and *APOBEC3B*). It appeared that the level of APOBEC enrichment correlated with the APOBEC expression score (**Figure 1E**). This analysis confirmed the link between APOBEC RNA expression at the time of biopsy and the historical accumulation of APOBEC-associated mutations in mUC that others have reported in primary urothelial carcinoma[15,16].

Recently, it was proposed that edited *DDOST* mRNA can be used to measure ongoing APOBEC mutagenesis[17]. We found that the frequency of C>U alterations in the *DDOST* mRNA at position

chr1:20981977 was enriched in tumors with APOBEC mutagenesis, with up to 15% of mRNA reads edited in one single sample (**Figure 1F**). Additionally, we analyzed ongoing APOBEC mutagenesis in mUC, by WGS of eight tumors from patients who had undergone serial biopsies of metastatic lesions (**Figure S4)**. We observed that the APOBEC mutational signature was present in private mutations of the second biopsy of these patients, suggesting that APOBEC mutagenesis could be active in the period between the first and second biopsy (**Figure S4A).** A lower cancer cell fraction (normalized allele frequency by copy number and purity, see Methods) in private SNVs of the second biopsy compared to shared (trunk) mutations confirms that these mutations were acquired more recently as they are only present in a subpopulation of cancer cells (**Figure S4B**). This result, together with the APOBEC signature detected in the subclonal mutations (**Figure S5**), suggests that the presence of subclonal populations due to APOBEC mutagenesis may contribute to the ongoing evolution of UC in the metastatic setting.

APOBEC-associated hotspot mutations are enriched in highly accessible genomic regions

The high resolution achieved by WGS, allowed us to investigate the enrichment of APOBEC and non-APOBEC mutations (non-TpC context) in specific genomic regions. We found that the number of non-APOBEC-associated SNVs, for instance, varied across the genome (**Figure 2A**). When this distribution overlapped with DNA accessibility and overall gene expression level, the frequency of non-APOBEC mutations decreased in open chromatin (highly accessible regions) and highly transcribed regions (**Figure 2B**). In contrast, the frequency of APOBEC-associated mutations was nearly constant across the genome. When restricting the analysis to APOBEC tetra-base mutations, we found that the difference in the distribution of Y/RTCA mutations across genomic regions decreases with the level of APOBEC enrichment (**Figure 2C**). Interestingly, in APOBEC-high tumors, YTCA mutations were evenly distributed. In contrast, RTCA mutations were enriched in low DNA accessible and low transcribed regions, although this enrichment was considerably less compared to tumors with lower levels of APOBEC mutations.

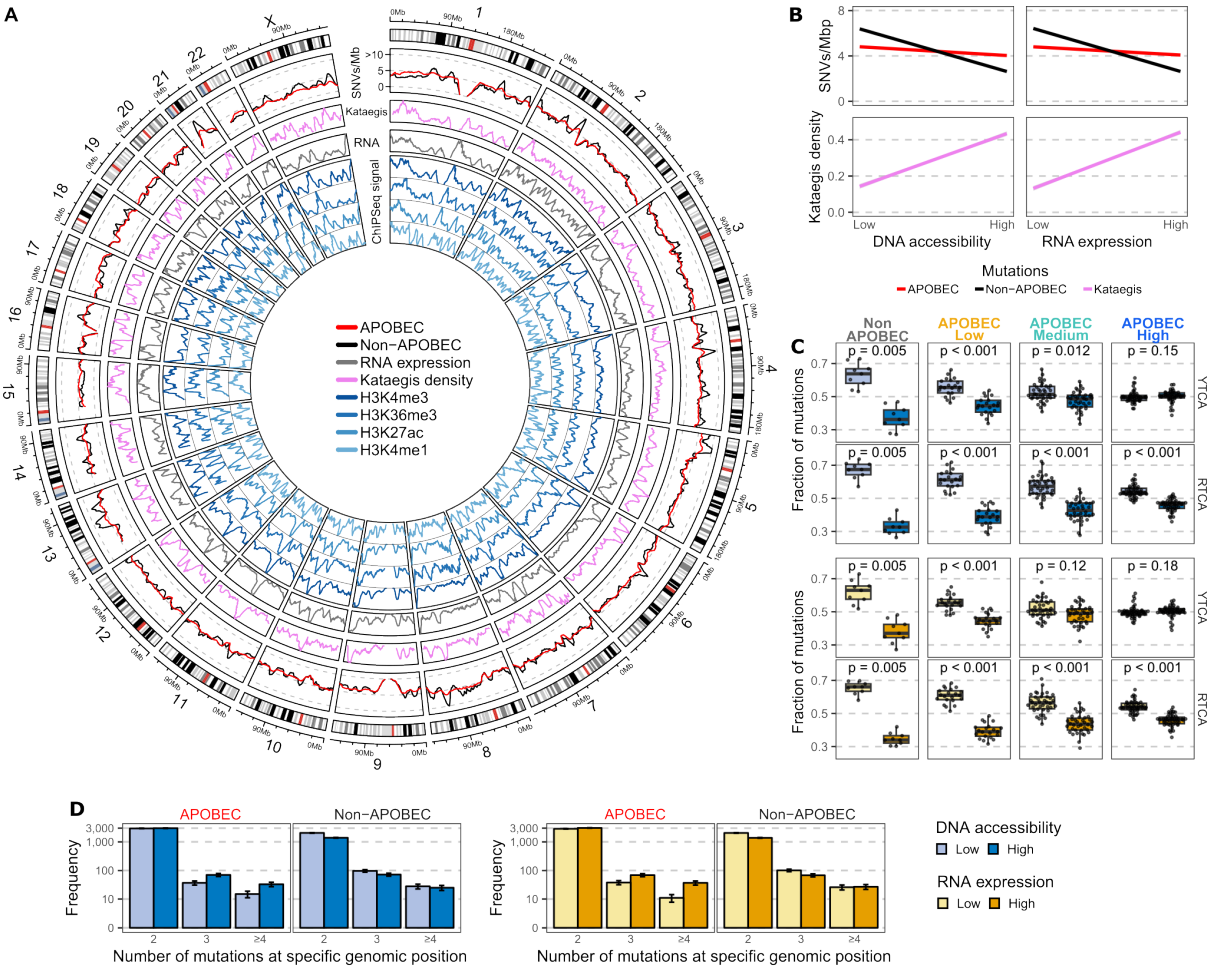


Figure 2. Distribution of APOBEC-associated mutations across genomic regions of metastatic urothelial carcinoma

(A) Whole-genome sequencing data (n = 115) were analyzed to estimate the mean number of single nucleotide variants (SNVs) in windows of one mega base pair (Mbp) across the entire genome. The circos plot shows from outer to inner circles: the genomics ideogram from chromosome 1 to X where the centrosomes are indicated in red; the mutational load of APOBEC and non-APOBEC associated mutations (mutations in TpC or non-TpC context, respectively); The density of kataegis events; Average RNA counts (expression) from tumors with available RNA-sequencing data (n = 90); DNA accessibility estimation from different

ChIPseq experiments of multiple histone marks from normal urothelial samples derived from ENCODE[45]. Peaks represent highly accessible DNA.

(B) Linear regression of the mutational load for APOBEC- and non-APOBEC-associated mutations as well as the density of kataegis events across the genome with DNA accessibility and expression data.

(C) Relative distribution of APOBEC YTCA and RTCA mutations across DNA-accessible and RNA expression regions. Samples were stratified according to the level of APOBEC mutagenesis. The Wilcoxon signed-rank test was applied and p-values were Benjamini-Hochberg corrected.

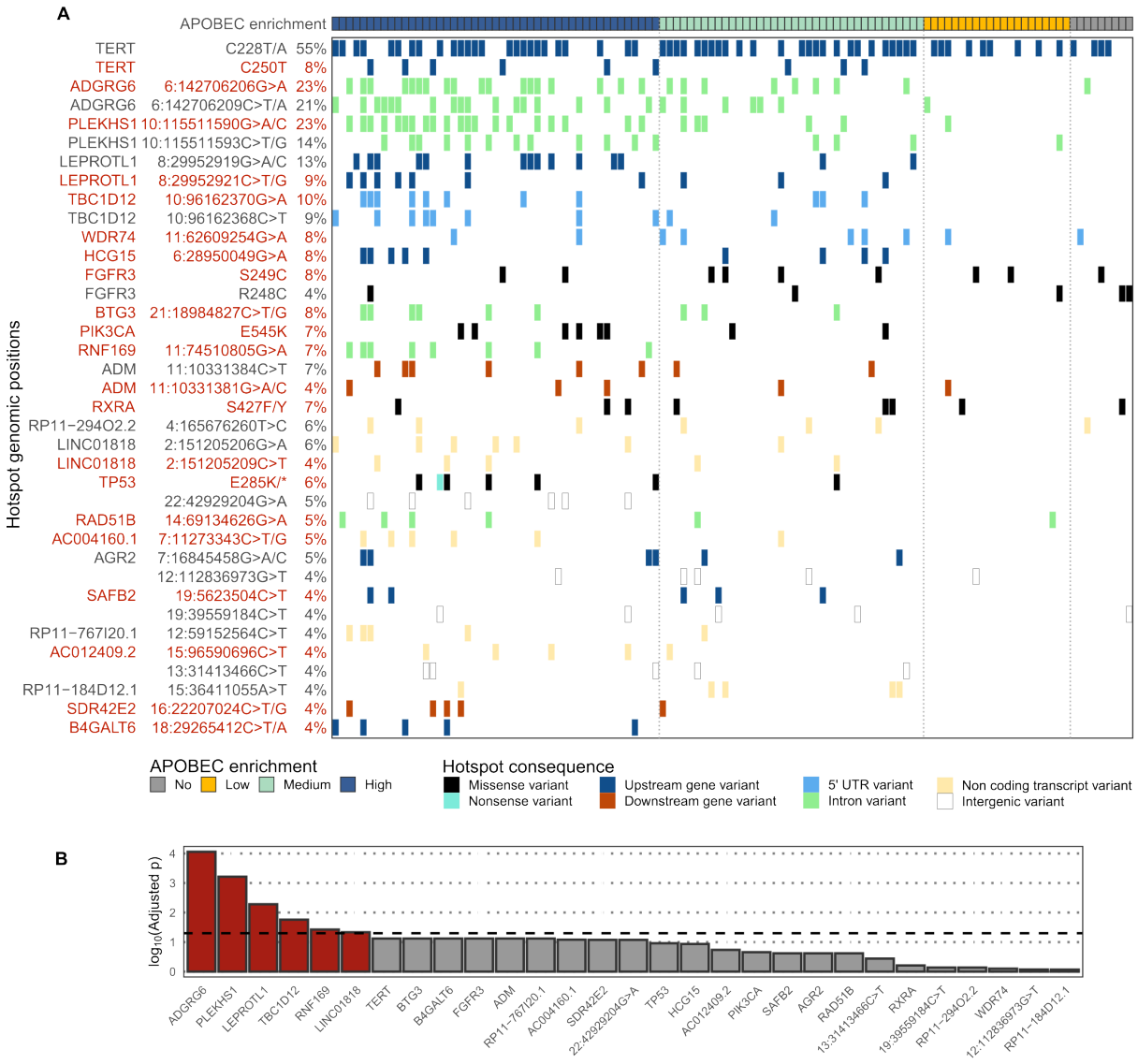
(D) Frequency of hotspot mutations grouped according to APOBEC and non-APOBEC-associated mutations, and DNA accessibility or RNA expression level.

Because of the high correlation between kataegis and APOBEC enrichment, we analyzed the distribution of kataegis loci across the genome. Contrary to the overall distribution of SNVs, our data suggest that kataegis events are more likely to occur in regions with high DNA accessibility and high transcriptional activity (**Figure 2B**). Moreover, we also evaluated the genome-wide distribution of all hotspot mutations (two or more mutations in a specific genomic position), representing 0.35% of all mutated genomic positions. We found that the frequency of highly recurrent (n≥4) ApoHM were enriched in high DNA accessibility and high transcriptional active regions (**Figure 2D**). Thus, while general APOBEC mutagenesis seemed to occur uniformly across the genome, kataegis and ApoHM seemed to occur more frequently in open chromatin and highly transcribed loci.

Recurrent hotspot mutations correlate with APOBEC mutagenesis

Next, we investigated the genomic consequence of hotspot mutations and found that the most frequent hotspot mutations in mUC occurred in non-coding regions of the genome (**Figure 3A**). Hotspot mutations in the *TERT* promoter were present in 62% of tumors. In line with previous reports[18,19], *TERT* expression did not differ between tumors with hotspot mutations and those being *wildtype* (**Figure S6A**). However, differential gene expression analysis showed that tumors with hotspot mutations in the *TERT* promoter had high expression of genes related to the biological oxidation pathway (**Figure S6B, Table S1**). Besides *TERT*, other frequent hotspot mutations were identified in the non-coding regions near *ADGRG6* (40%), *PLEKHS1* (33%), *LEPROTL1* (20%), and *TBC1D12* (15%). Similarly, these hotspot

209 mutations did not affect the expression of these genes but were associated with transcriptomic effects
210 in several genes (**Figure S6)** and biological pathways (**Table S1**). These hotspot mutations strongly
211 correlated with enrichment for APOBEC-associated mutations, suggesting their origin in APOBEC
212 mutagenesis (**Figure 3B**).



213

214 **Figure 3. Recurrent hotspot mutations of metastatic urothelial carcinoma correlate with APOBEC**
215 **mutagenesis**

216 (A) Overview of recurrent hotspot mutations present in at least five samples, including all
217 substitutions occurring in the same genomic position. Hotspot mutations occurring in the TpC
218 context are highlighted in red.

219 (B) The association of hotspot mutations and APOBEC fold enrichment (continuous values) was
220 interrogated with a logistic regression analysis applying the Wald test. P-values were

corrected using the Benjamini-Hochberg method and ordered accordingly. Bars above the dashed line ($-\log_{10}(0.05)$) are statistically significant and are indicated in red. See also Figure S6 and Table S1.

All frequent hotspot mutations in the coding region have been previously described and affected known driver genes: *FGFR3* S249C/R248C (8%, 4%), *PIK3CA* E54K (7%), *RXRA* S427F/Y (7%) and *TP53* E285K/* (6%). Comparing the expression of these known driver genes affected by hotspot mutations *vs.* the wildtype, only *FGFR3* hotspot mutations significantly affected the expression of this gene (**Figure S6A**).

Hairpin-loops are targets of twin hotspot mutations called Didymi

We noticed that the hotpot mutations near *ADGRG6*, *PLEKHS1*, *LEPROTL1* and *TBC1D12* are located within DNA hairpin-loop structures. It is known that DNA hairpin-loops are targets of APOBEC3A (**Figure 4A**)[20,21], therefore, we predicted DNA hairpin-loops for all mutated genomic positions (see Methods). The predicted hairpin-loops near *ADGRG6*, *PLEKHS1*, *LEPROTL1* and *TBC1D12* are each affected by two hotspot mutations, which are referred to as twin mutations[21]. Moreover, we noticed that the twin mutations were not mutually exclusive, which differs from the hotspot mutations in the *TERT* promoter (mutual exclusivity test, $p < 0.001$). Only the twin hotspot mutations in *TBC1D12* co-occurred more frequently than expected among APOBEC-high tumors ($p = 0.02$). Further analysis of co-occurred twin hotspot mutations revealed that very few had identical variant allele frequencies, suggesting that most twin mutations in the same tumor occurred in independent events as they were also found on different alleles (**Figure S7**).

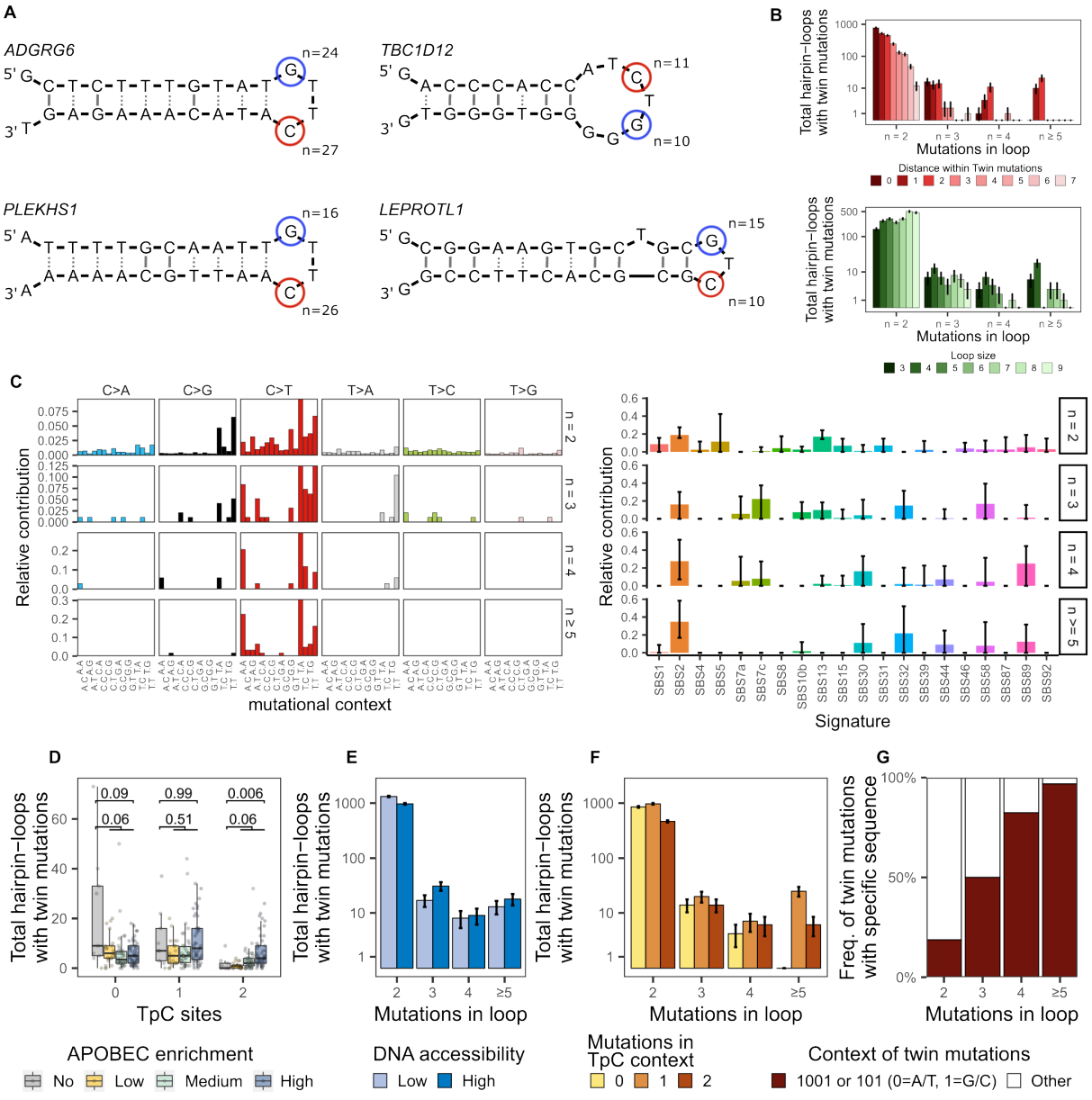


Figure 4. Genomic characteristics of twin hotspot mutations in metastatic urothelial carcinoma

(A) Hairpin-loop structures affected by frequent hotspot mutations in *ADGRG6*, *TBC1D12*, *PLEKHS1* and *LEPROTL1*. The positions of hotspot mutations are marked in red for TpC context and blue otherwise.

(B) Distribution of twin mutations according to the distance between twin mutations and loop size.

(C) Mutational signatures (COSMIC v3.3) of twin mutations according to their frequency. The stability of the signature call was tested by applying 1,000 bootstrap iterations. Only SBS2 was very stable in highly frequent (n ≥ 5) twin mutations.

(D) Frequency distribution of hairpin loops affected by twin mutations according to APOBEC mutagenesis,

(E) DNA accessibility,

(F) number of mutations in TpC context within a loop and

(G) DNA sequence between twin mutations. Wilcoxon rank-sum test was applied to compare APOBEC *vs* non-APOBEC tumors and p-values were Benjamini-Hochberg corrected. See also Figures S7-S9 and Tables S2 and S3.

Next, we investigated the properties and origin of these twin mutations. A comprehensive analysis of all DNA hairpin-loop structures in the human genome affected by two mutations in their loops revealed 2,387 twin mutations (4,774 altered genomic positions), representing 0.16% of all mutated genomic positions. The distance between twin mutations varied, but when the frequency of mutations increased, the distance decreased to mainly one or two bases and the loop sizes to mainly three to four bases (**Figure 4B**). Additional examination of the 96 tri-nucleotide contexts of all twin mutations revealed that both APOBEC COSMIC signatures, SBS2 and SBS13, were dominant (**Figure 4C**). However, at higher mutational frequencies ($n \geq 5$), only signature SBS2 remained. We also observed a secondary signature of C>T mutations in the ApC context that does not resemble any known COSMIC signature (**Figure S8, Table S2**). The absolute contribution of this signature was similar across all APOBEC tumors and its prevalence in the mUC cohort correlated with spontaneous deamination (SBS1) and defective DNA mismatch repair signatures (**Figure S9**; SBS6, SBS15, SBS20 and SBS26).

Furthermore, APOBEC-driven tumors were enriched for twin mutations occurring only in the TpC context (**Figure 4D**). Contrary to the general pattern of ApoHM (enriched in DNA accessible regions), twin mutations with a high number of alterations were similarly distributed between high and low DNA accessible regions (**Figure 4E**).

Additionally, we found that at higher mutational frequency, at least one of the twin mutations occurs in the TpC context (**Figure 4F**) and the sequence between the two is 1001 or 101 (0 = A/T, 1 = G/C; underlined are the positions of the twin mutations) (**Figure 4G**). Because of the unique characteristics of frequently affected twin mutations, we named them didymi (twins in Greek). In summary, didymi are two C>T hotspot mutations found in DNA hairpin-loops separated by one or two A/T base-pairs in which at least one of the twin mutations is located in TpC context and the other in NpC (N = any base-pair;

most N bases are A or T). Applying this definition, we identified 231 didymi in the mUC cohort (**Table S3**).

Driver hotspot mutations in urothelial carcinoma

After identifying several hotspot mutations that could be attributed to APOBEC activity, we assessed whether these hotspot mutations had a selective advantage (drivers) or not (passengers). Recent attempts relying on the stability of hairpin-loops have been proposed to differentiate passengers from driver ApoHM[8,20]. We confirmed that a more stable loop (Gibbs free energy ΔG; see STAR Methods) leads to a higher number of alterations (**Figure 5A**). Taking this into account, we developed a statistical model to identify driver hotspot mutations considering not only the stability of hairpin-loops but also the tri-nucleotide context, DNA accessibility and the potential for didymi via sequence in the loop (**Figure 5A**).

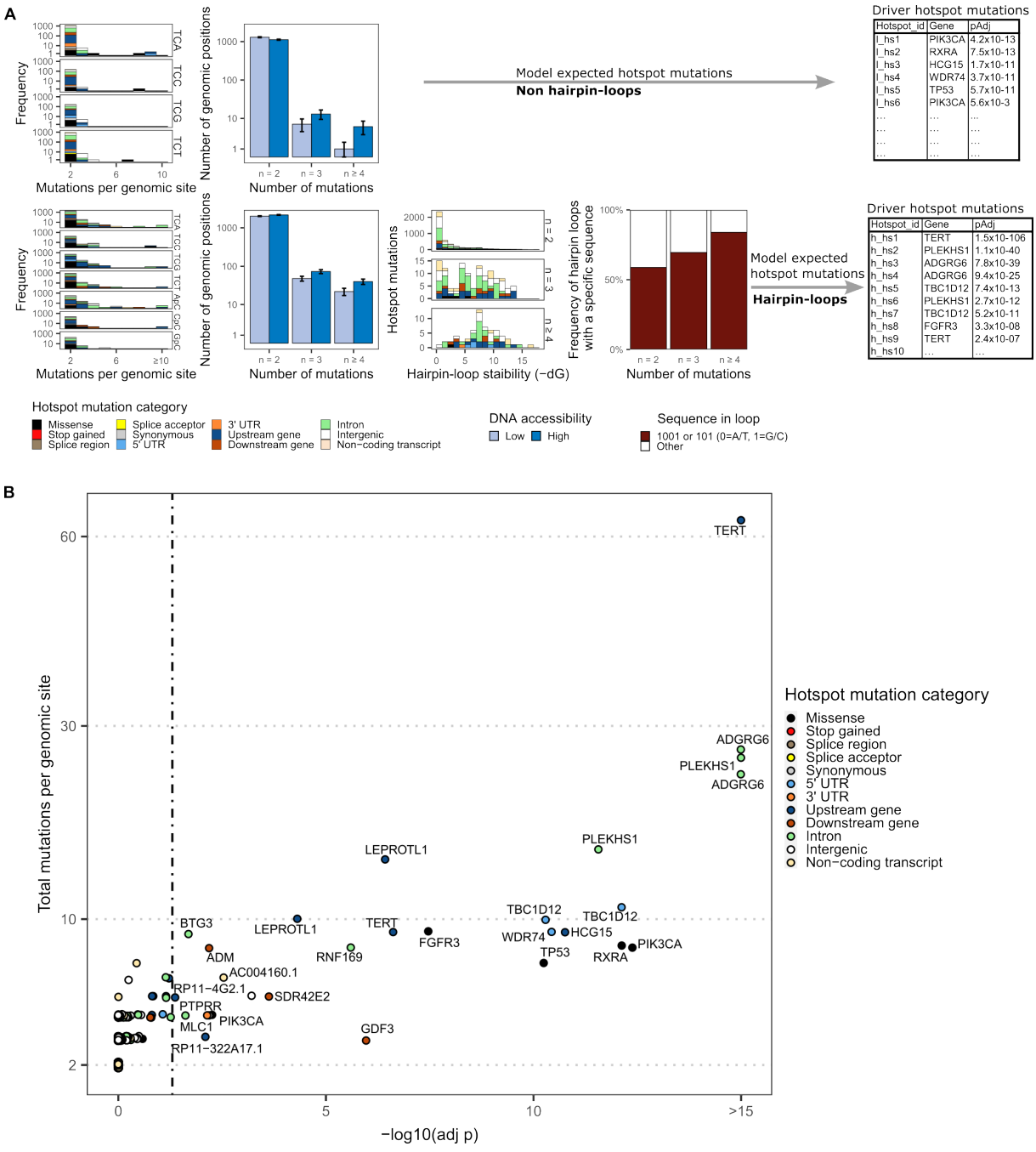


Figure 5. Driver hotspot mutations associated with APOBEC mutagenesis in urothelial carcinoma

(A) Frequency distribution of variables (trinucleotide context, DNA accessibility, DNA hairpin-loop stability and sequence in the loop) considered to identify hotspot mutations that were more frequently mutated than expected outside (top) and within (bottom) hairpin-loop structures.

(B) Driver hotpot mutations were estimated separately for mutations outside and within hairpin loops. Per group, p-values were corrected using the Benjamini-Hochberg method.

See also Figures S10 and S11, and Table S4.

303

304 Putative ApoHM were divided into those located outside or inside the loop of DNA hairpin-loop
305 structures. In case of ApoHM outside hairpin-loops, only those in the TpC context were considered. For
306 ApoHM within loops, all alterations in TpC, ApC, CpC and GpC context were included in the analysis to
307 account for didymi. In case of TpC, the distribution of hotspot mutations in the trinucleotide context was
308 considered. A background distribution was modeled as a Poisson process and the significant enrichment
309 of mutations in a particular genomic site was estimated.

310 We identified 0.40% (n = 27) of ApoHM as drivers (adjusted p < 0.05; **Fig 5B**). Known driver genes
311 affected by hotspot mutations included coding alterations in *TP53, PIK3CA, FGFR3, RXRA* and the
312 *TERT* promoter. All other putative driver ApoHM affected non-coding regions including didymi in
313 *ADGRG6*, *PLEKHS1*, *TBC1D12* and *LEPROTL1* proposed as drivers by other studies[9,22]. Other
314 potential driver ApoHM include *RNF169* (involved in DNA damage repair)[23], *BTG3* (angiogenesis)[24],
315 *ADM* (adrenomedullin; vasodilator)[25], *GDF3* (regulation of TGF-beta)[26] and WDR74 (ribosome
316 biogenesis)[27].

317 To validate our method and confirm the driver ApoHM assessment, we used an independent cohort of
318 non-metastatic urothelial carcinoma of the bladder (n = 23) of the pan-cancer analysis of whole-genomes
319 (PCAWG) study[28] (**Figure S10**). This analysis confirmed the previously identified ApoHM as potential
320 cancer drivers of urothelial carcinoma. Moreover, in this cohort, 96% of tumors were APOBEC-driven,
321 APOBEC enrichment correlated with kataegis and the largest group (35%) had high enrichment for
322 APOBEC-associated mutations.

323 Furthermore, the performance of the model to identify driver ApoHM in hairpin loops was tested. The Q-
324 Q plots show that the empirical distribution of ApoHM deviates from the theoretically expected
325 distribution (Kolmogorov–Smirnov test p < 0.001). However, when outliers that represent highly frequent
326 ApoHM (>10) are excluded, which according to our analysis are all drivers, we observed a good
327 agreement between our model and the theoretical distribution (**Figure S11A**; Kolmogorov–Smirnov test
328 p = 0.19). By simulating a synthetic genome of mUC, we showed that an 80% statistical power is reached
329 when the cohort size is ~75 samples for highly frequent (>10%) driver ApoHM (**Figure S11B**). However,
330 the power to detect rare driver ApoHM (≤10%) is considerably reduced, and a larger cohort is needed.
331 We also evaluated the contribution of different genomic features as covariates (**Table S4**) to identify

driver ApoHM. The McFadden's R^2 in the model that only considers the trinucleotide context is low (R^2 = 0.04), but the goodness of fit increases when considering the hairpin loop (R^2 = 0.23), hairpin-loop + sequence in the loop (R^2 = 0.27) and when adding DNA accessibility regions (R^2 = 0.28) into the model. DNA accessibility shows a high (anti-)correlation with other genomic features: GC content, RNA expression levels, mutational load, methylation and replication timing (**Figure S11C**). Therefore, the addition of the other variables in the model has limited added value (**Figure S11D**).

Driver hotspot mutations in metastatic breast cancer

To test our statistical framework in other cancer types and to evaluate if our findings were UC-specific, we analyzed a cohort of 442 mBC (**Figure 6**)[29]. Similar to UC, breast cancer is commonly affected by APOBEC mutagenic activity[1]. We identified APOBEC-enriched tumors in 76% of patients, and only 19% of mBC tumors were APOBEC-high. In most patients (39%), tumors were classified as APOBEC-low and were enriched for HR deficiency (two-sided Fisher's exact $p < 0.001$).

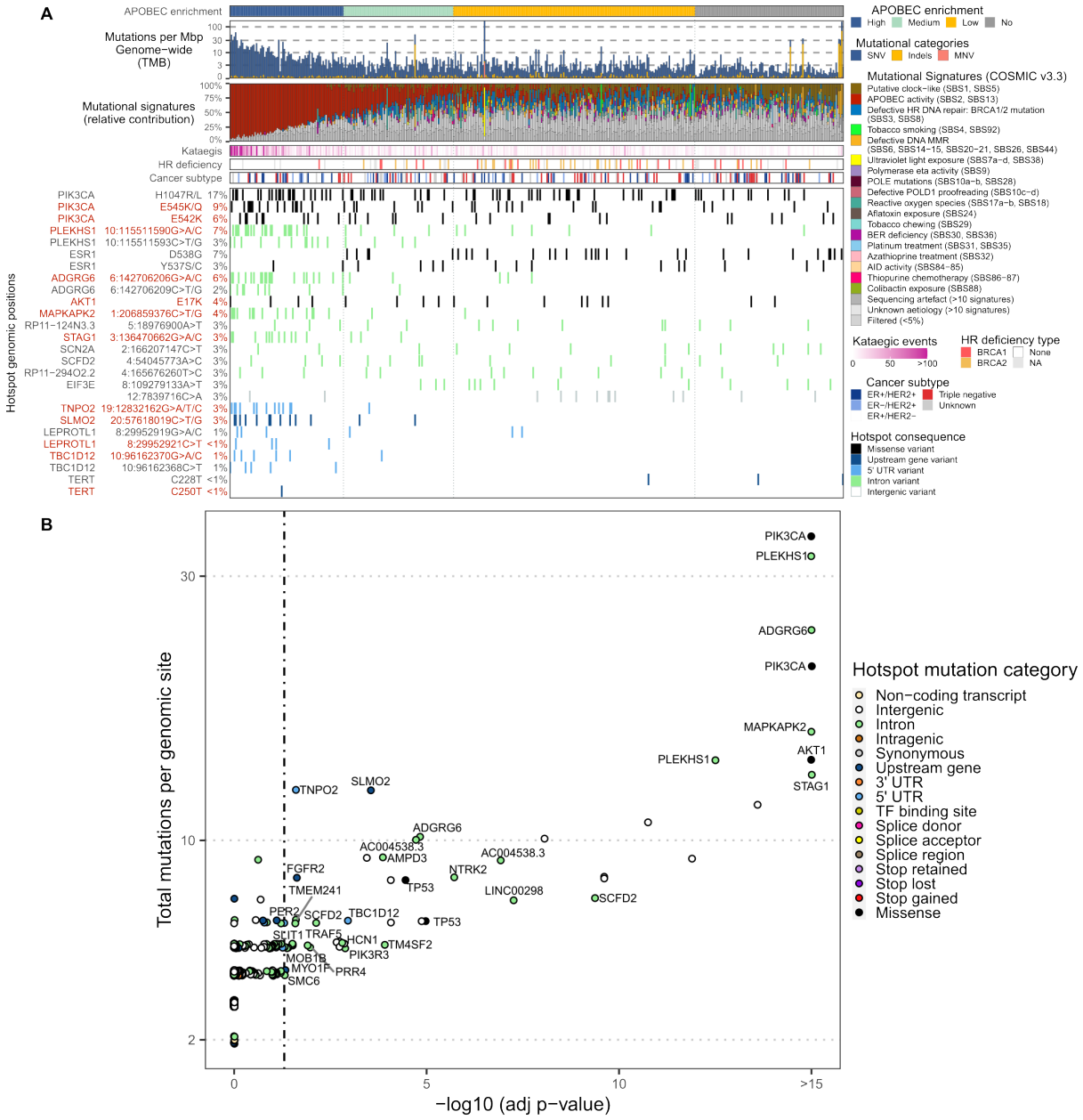


Figure 6. Genomic landscape and driver hotspot mutations associated with APOBEC mutagenesis in metastatic breast cancer

(A) Whole-genome sequencing data from 442 metastatic breast cancer were analyzed and patients were classified according to the enrichment of APOBEC-associated mutations. The genomic features are displayed from top to bottom as follows: APOBEC mutagenesis; genome-wide tumor mutational burden (TMB); COSMIC mutational signatures; frequency of kataegis; homologous recombination (HR) deficiency and its probable origin; cancer subtype and; the most frequent hotspot mutations (known hotspot mutations in *LEPROTL1*, *TBC1D12* and *TERT* were also included).

(B) Putative driver hotspot mutations in APOBEC-enriched breast cancer. P-values were adjusted using the Benjamini-Hochberg method.
See also Figure S12 and Table S5.

The most frequent coding hotspot mutations affected *PIK3CA*, *ESR1* and *AKT1*. Twin mutations in hairpin-loops displayed a similar mutational signature as those in mUC, including the APOBEC signature SBS2 in conjunction with the uncharacterized C>T mutations in ApC context that define didymi (**Figure S12A-B**). Didymi in *PLEKHS1* and *ADGRG6* were the most frequent non-coding hotspot mutations, while *LEPROTL1* and *TBC1D12*, two other didymi frequently found in mUC, only affected ≤1% of mBC. A total of 694 didymi were identified in mBC (**Table S5**), but only 19 (2.7%) were shared with mUC (**Figure S12C**).

Our analysis revealed 51 driver ApoHM in APOBEC-enriched mBC (**Figure 6B**), representing only 0.07% of all ApoHM. Drivers included missense hotspot mutations in *PIK3CA*, *AKT1* and *TP53*, and hotspot mutations outside of the protein-coding region of *MAPKAPK2*, *STAG1* and including didymi in *PLEKHS1* and *ADGRG6*. In contrast to mUC, and despite being one of the most frequently affected genes by hotspot mutations in mUC, *LEPROTL1* was not a driver of mBC. This analysis suggests that driver mutations derived from APOBEC activity are cancer-type-specific.

Driver hotspot mutations across multiple metastatic cancers

APOBEC mutagenic activity is widespread across multiple cancer types. Here, we analyzed the genome of 16 additional metastatic cancers, which in total represents 3302 whole genomes (+115 mUC +442 mBC = 3859). Urothelial, breast and uterus cancers have the highest proportion of APOBEC-high tumors (**Figure 7A**). In mUC and mBC, 95% of hotspot mutations affect non-coding transcripts, introns or intergenic regions. This proportion varies per cancer type and can represent up to 99% of all hotspot mutations in esophagus cancer. Missense hotspot mutations are rare, but the highest proportion is found in liver cancer, reaching 5%.

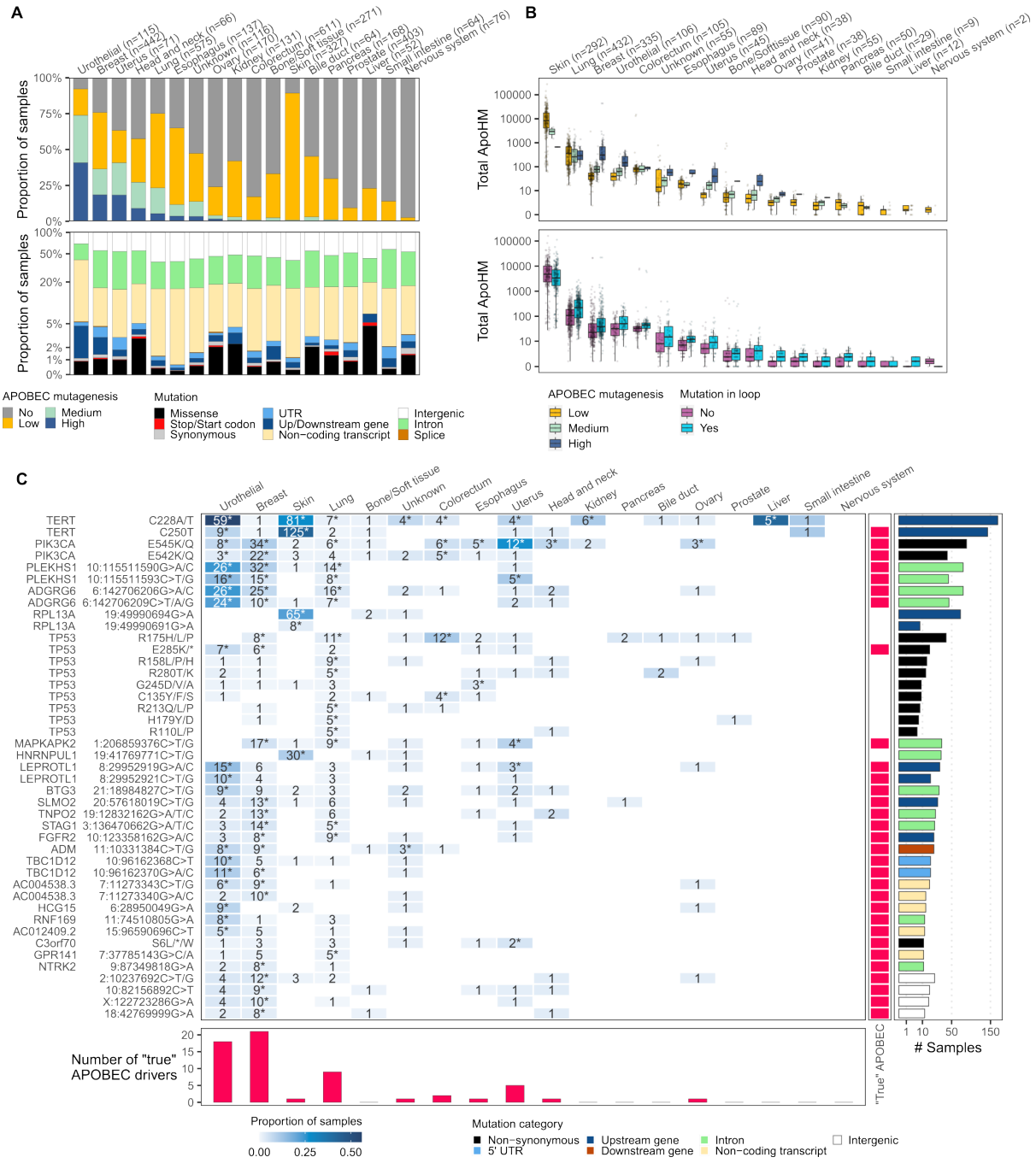


Figure 7. Pan-cancer overview of APOBEC-derived hotspot mutations and drivers.

(A) Proportion of APOBEC-enriched tumors and hotspot mutations across cancer types.

(B) Distribution of APOBEC-associated hotspot mutations (ApoHM) across the genome of APOBEC-enriched tumors.

(C) The 10% most frequent genomic positions or genes affected by driver ApoHM. Drivers of specific cancer type are indicated by stars. The association of driver ApoHM with APOBEC fold enrichment (continuous values and using all tumor samples) using logistic regression analysis

and applying the Wald test (p adj < 0.05) shows the "true" APOBEC-derived mutations. All "true" APOBEC-derived driver ApoHM are included in the figure.
See also Tables S6 and S7.

The frequency of ApoHM increases with the strength of APOBEC mutagenesis and are more recurrent in hairpin loops (**Figure 7B**). However, skin cancer does not follow this pattern, which has been proven to be problematic in other studies due to its hypermutated nature, inflating the number of driver events[30,31]. These studies make special considerations or exclude skin cancer altogether from their analysis. We found that skin cancer is mostly defined as an APOBEC-low cancer type and we suspected that a large proportion of mutations that have the APOBEC signature may not have derived from APOBEC mutagenic activity. This is supported by the relatively low number of driver ApoHM in skin cancer that correlates with APOBEC enrichment (**Figure 7C**), despite many ApoHM defined as drivers by our model (**Table S6**). These driver ApoHM considered "true" APOBEC-derived mutations are more frequent in breast, urothelial, lung and uterus cancers. *TERT*, *PIK3CA*, *PLEKHS1* and *ADGRG6* are the most affected genes by driver ApoHM. All four genes harbor two driver ApoHM, which are targeted by APOBEC except for *TERT* that has only one "true" APOBEC-derived hotspot mutation (C250T). *TP53* is another gene that is frequently affected by driver ApoHM, however, only one out of nine is a "true" APOBEC-derived mutation. The pan-cancer overview exhibits the distribution of driver ApoHM in APOBEC-enriched cancers, and the power gained by integrating 3859 samples revealed that of all potential drivers, only 31 might be "true" APOBEC-derived mutations.

DISCUSSION

In this study, we describe the genomic landscape of APOBEC-driven tumors, characterize ApoHM and identify potential cancer drivers in mUC. The in-depth analysis of 115 whole-genomes of mUC identified chromatin accessibility, hairpin-loop stability and specific sequences within the hairpin-loop as variables associated with ApoHM. These variables, in combination with the mutational context, were used to identify ApoHM that were more frequently mutated than expected by chance.

The substrate of APOBEC enzymes is single-stranded DNA (ssDNA)[1], which has led to the following conflicting hypotheses: 1) APOBEC enzymes are mainly active during replication[32] and 2) APOBEC is mainly active in open chromatin and transcriptionally active genomic regions[33]. The equal distribution of all APOBEC-associated mutations across genomic regions supports the hypothesis that these mutations had been generated during replication when APOBEC enzymes have equal access to ssDNA across the genome[32]. However, kataegis which has previously been linked to APOBEC activity[34], and ApoHM were enriched in high DNA accessible and highly transcribed regions. This observation reconciliates both views claiming that APOBEC is active during DNA replication (non-clustered and non-hotspot mutations) and transcriptionally active regions (clustered and hotspot mutations). Additionally, we observed that APOBEC3A-preferred YTCA mutations are dominant in mUC and are evenly distributed across genomic regions. This result is in line with experimental observations in human cancer cell lines[34], suggesting that APOBEC3A is the main driver of APOBEC mutagenesis. For YTCA mutations as well as for APOBEC3B-preferred RTCA mutations, the gap between the number of mutations across genomic regions is smaller at higher APOBEC mutagenesis, which strongly suggests that APOBEC enzymes do not have a preference for specific genomic regions. However, the enrichment of highly frequent ApoHM (n≥4) in open chromatin, of which many are drivers according to our analysis, may imply a functional effect of these putative driver mutations occurring near gene regulatory elements[22].

The extensive examination of ApoHM in hairpin-loops revealed twin mutations we termed didymi that are characterized by a unique mutational pattern. Didymi comprises the APOBEC SBS2 signature and an unknown signature delineated by C>T mutations in the ApC context. It is remarkable to see only one of the APOBEC signatures in didymi loci, when they usually appear together in tumor samples with APOBEC mutagenesis[35,36]. The fact that only C>T mutations that characterize SBS2 are present in didymi suggests that these mutations may arise predominantly by replication across the uracil

bases[37,38] and that the mechanisms to generate C>G mutations that characterize SBS13 are not operational in this context. Furthermore, the unknown mutational signature of didymi correlates with spontaneous deamination (SBS1) and defective DNA mismatch repair signatures (SBS6, SBS15, SBS20 and SBS26), suggesting a potentially different mechanism in ApC from the putative TpC APOBEC mutations. Although there is a strong correlation with APOBEC mutagenesis[21], it is unclear if both mutations in didymi loci are direct targets of APOBEC3A or whether the non-TpC mutations are just the result of spontaneous deamination followed by DNA mismatch repair.

Compared to breast cancer, twice as many bladder cancer tumors were APOBEC-high (19% vs. 41%). Most driver ApoHM were cancer-specific, possibly reflecting different selective pressures that each cancer type endured. In both tumor types, APOBEC-low patients had enrichment for HR deficiency, while APOBEC-high tumors had high tumor mutational burden which may indicate different treatment options for these two groups of patients with different levels of APOBEC mutagenesis[39-42].

The in-depth analysis performed in this study to characterize the mutational landscape of APOBEC mutagenesis, revealed the correlation of ApoHM with the stability of DNA hairpin-loops, DNA accessibility and the potential for didymi associated with these recurrent mutations. These features are key to modeling the background distribution of hotspot mutations and identifying potential drivers. Most potential driver ApoHM were in the non-protein coding regions, including didymi. The similar frequency of these drivers in the metastatic and primary settings of UC indicates a general phenomenon in UC and the drivers could cause early events of tumorigenesis of UC. APOBEC-associated mutations have also been identified in normal tissue[43,44], however, it is unclear if the driver ApoHM we report here are also present in healthy tissue and to what extent they contribute to cancer development from normal cells. Nevertheless, experimental validation will be needed to confirm the cancer driver status of these hotspot mutations.

Although several hotspot mutations are defined as drivers by our model, the inclusion of other cancer types, increased the statistical power revealing that only 31 driver ApoHM have a strong correlation with APOBEC, and may be considered “true” APOBEC-derived hotspot mutations.

In this study, we characterized the genomic landscape of APOBEC-driven mUC and identified novel mechanisms of genomic alteration patterns associated with APOBEC mutagenesis. The mutational signatures associated with DNA hairpin-loops targeted by APOBEC in two distinct hotspot positions are

unique, demonstrating the exclusive mutational signature of APOBEC-derived hotspot mutations. These findings were confirmed in non-metastatic UC and in metastatic breast cancer. Additional studies are needed to clarify the role of APOBEC in these recurrent twin mutations. Also, the enrichment of ApoHM and kataegis in high DNA accessible regions, suggests a different mechanism compared to the general APOBEC mutagenesis (non-hotspot mutations) that seems to occur independently of genomic regions, which may be linked to different mechanisms of APOBEC3A and APOBEC3B[13]. As APOBEC is a major source of hotspot mutations, it is crucial to identify those in coding and non-coding regions of whole-genomes that may play an important role in cancer development. The statistical framework we developed could aid to identify potential driver hotspot mutations derived from APOBEC activity which may offer novel targeted therapy options for APOBEC-driven cancer patients.

Limitations of the study

Despite the thorough analysis we have performed, caution should be exercised when considering these outliers as true APOBEC-derived driver hotspot mutations as other unknown factors may still explain the distribution of these highly frequent APOBEC-related hotspot mutations. The sample size for some tumor types is a limitation when identifying driver hotspot mutations in a cancer-specific manner as we did.

ACKNOWLEDGEMENTS

Hartwig Medical Foundation and the Center of Personalized Cancer Treatment are acknowledged for making the clinical, genomic and transcriptomic data available We would also like to thank the Pan-Cancer Analysis of Whole Genomes study for sharing the genomic data from 23 primary UC patients. We are particularly grateful to all participating patients and their families. The Stichting Dutch Uro-Oncology Study (DUOS) group and the Daniel den Hoed Foundation supported this research. J. Alberto Nakauma-González was collectively supported by the national funding organization the Dutch Cancer Society (KWF, the Netherlands) under the framework of the ERA-NET TRANSCAN-2 initiative.

AUTHORS CONTRIBUTIONS

Conceptualization: JAN and HJGvdW; Methodology: JAN and HJGvdW; Software: JAN, HJGvdW and MTWN; Validation: JAN and HJGvdW; Formal Analysis: JAN and HJGvdW; Investigation: All authors; Resources: JLB, JWMM; Data Curation: JAN, MR and HJGvdW; Writing – Original Draft: JAN and HJGvdW; Writing – Review & Editing: All authors; Visualization: JAN; Supervision: JLB, HJGvdW, MPJL and JWMM; Project Administration: JAN, HJGvdW, MPJL and JLB; Funding Acquisition: JLB , HJGvdW, and MPJL.

DECLARATION OF INTERESTS

Joost L. Boormans has received research support from Merck AG / Pfizer, Janssen and Merck Sharp & Dohme, and consultancy fees from Merck Sharp & Dohme, Bristol-Myers Squibb, Astellas, AstraZeneca, Ipsen and Janssen (all paid to the Erasmus MC Cancer Institute).

Martijn P. J. Lolkema has received research support from JnJ, Sanofi, Astellas and MSD, and consultancy fees from Incyte, Amgen, JnJ, Bayer, Servier, Roche, INCa, Pfizer, Sanofi, Astellas, AstraZeneca, Merck Sharp & Dohme, Novartis, Julius Clinical and the Hartwig Medical Foundation (all paid to the Erasmus MC Cancer Institute).

John W. M. Martens has received research support from Pfizer, Sanofi, GSK, Therawis Cergentis, and Philips (all paid to the Erasmus MC Cancer Institute) and one consultancy fee from Novartis.

Astrid A.M. van der Veldt has received consultancy fees from BMS, MSD, Pfizer, Novartis, Eisai, Sanofi, Pierre Fabre, Ipsen and Roche (all paid to the Erasmus MC Cancer Institute).

516 J. Alberto Nakauma-González, Maud Rijnders, Minouk T.W. Noordsij and Harmen J. G. van de Werken

517 declare no competing interests.

518

STAR METHODS

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Harmen J. G. van de Werken (h.vandewerken@erasmusmc.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

WGS, RNA-seq and clinical data from mUC, mBC and from other metastatic cancers are available through the Hartwig Medical Foundation at https://www.hartwigmedicalfoundation.nl, under request numbers DR-314, DR-026 and DR-041 respectively. For mUC, samples that were previously analyzed (DR-031) by Nakauma et al.[10] were retrieved from DR-314. WGS data from primary UC was requested to the NCBI dbGAP and granted access through request #33427.

ChIPseq, replication timing and methylation data experiments, are freely available through The ENCODE Project Consortium[46] and the Roadmap Epigenomics Consortium[47] on the ENCODE portal (https://www.encodeproject.org)[45].

The scripts, including the algorithm to find hairpin-loops and estimate the thermodynamic stability have been deposited in a public repository available at https://github.com/ANakauma/ApobecHM_drivers. Additionally, the version v1.0.0 of the code used for this study (ApobecHM_drivers) is available at Zenodo (https://doi.org/10.5281/zenodo.10362579)[48]. Pre-processed WGS data was provided by the Hartwig Medical Foundation and scripts are available at https://github.com/hartwigmedical/hmftools. R2CPCT v0.4 was used for additional processing of the WGS (https://github.com/J0bbie/R2CPCT),

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patient cohorts

The mUC cohort of this study has been previously described (NCT01855477 and NCT02925234)[10]. In short, patients with advanced or mUC were prospectively enrolled in these multicenter clinical trials and

were scheduled for 1st or 2nd line palliative systemic treatment. Following protocols of the Hartwig Medical Foundation (HMF)[49], WGS, with a depth close to 100X[10], was successfully performed on DNA from freshly obtained biopsies from metastatic sites, and matched RNA-sequencing (RNA-seq) was available for 90 patients (97 samples). Sequential biopsies of a metastatic lesion taken at the time of clinical or radiological disease progression from eight patients were additionally sequenced. Similarly, the cohorts of other cancer types have been previously described[29], and the DNA extraction and sequencing were performed according to the HMF protocols[49]. Only cancer types with >50 samples were included in the analysis.

METHOD DETAILS

Whole-genome sequencing and analysis

Alignment and pre-processing of WGS data, detection of genomic subtypes, HR deficiency, MSI, structural variants, chromothripsis events, APOBEC mutagenesis, and pathway activity have been previously described[4,10,49,50]. Mutational signatures and kataegis were detected with MutationalPatterns v3.10.0[51] and Katdetectr v1.2.0[52]. APOBEC enriched tumors (adj. $p < 0.01$, otherwise non-APOBEC tumors) were classified as high when the fold enrichment (E) for C>T and C>G mutations in TCW (W = A or T) context was $E \geq 3$, medium when $2 \leq E < 3$ and low when $E < 2$. Similarly, the fold enrichment for C>T and C>G mutations in the tetra YTCA (Y = T or C) and RTCA (R = G or A) context was calculated.

Clonal fraction and cancer cell fraction

The clonal fraction of mutations was estimated as previously described[53]. Correcting for tumor purity and copy number, the variant copy number n_{SNV} of each SNV was calculated as follows

$$n_{SNV} = \frac{f_m}{p}[pC_t + (1-p)C_h], \quad (1)$$

where f_m is the relative frequency of the mutant variant reads, p is the tumor purity, C_t is the copy number affecting the region where a particular SNV was located and C_h is the healthy copy number (2 for autosomes and 1 for allosomes). In this study, mutations were considered clonal when the variant copy number was >0.75.

To identify the proportion of cancer cells carrying a specific mutation, the cancer cell fraction (CCF) was estimated as previously described[54,55]. Given the number of reference and mutant reads and assuming binomial distribution, we estimated the expected number of allelic copies (n_{chr}) carrying the observed SNV resulting from f_m values when the mutation is present in 1, 2, 3,…, N_{chr} allelic copies. The resulting estimated n_{chr} with the maximum likelihood is used to calculate the CCF as n_{SNV}/n_{chr}.

Mutational load across genomic regions

The genome was divided into regions (bins) of one mega base-pair (Mbp). The number of SNVs was counted in each bin, and the mean number of SNVs was estimated from the entire cohort. These values represented the average SNVs/Mbp reflecting the mutational load in each genomic region. The average SNVs/Mbp was smoothed by applying a moving average with a window of k = 3. For visualization reasons, in Figure 2 a k = 9 was used.

DNA accessibility estimation (ChIPseq)

ChIPseq data for healthy urinary bladder, breast and other tissues of adult humans (H3K4me1, H3K4me3, H3K36me3 and H3K27ac) were downloaded from the ENCODE portal (https://www.encodeproject.org) to our local server. The bed.gz files were imported with narrowPeak format for analysis. Only peaks with q < 0.05 were kept for analysis. The signal of each experiment was divided into regions of one Mbp, and a moving average with k = 3 bins was applied. The signals were normalized using the mean and standard deviation. This procedure was applied to each chromosome. The sum of all four ChIPseq experiments was considered an approximation of DNA accessibility. High DNA accessible regions (open chromatin) had values above the median considering the whole genome. All other regions were considered to be of low DNA accessibility (condensed chromatin). DNA accessibility for all healthy tissues is available in **Table S7** (for urinary bladder see **Table S4** along with other covariates). In case that matched normal ChIPseq with tumor type was not available, an average of all ChIPseq experiments was used.

Detection of hairpin loops

All SNVs were assessed to determine whether they occur in the loop of hairpin-loop structures and their thermodynamic stability. A total of 50 bases upstream and downstream of the mutation site were considered. The minimum length of the stem was 2 base-pairs and the minimum and maximum loop size was 3 and 10 bases, respectively (not considering the closing base-pair). Since multiple configurations are possible, only the structure with the highest stability was considered. One mismatch was allowed which could be either a non-matching base-pair or a single nucleotide bulge loop.

Stability of hairpin loops

We implemented the nearest neighbor stability algorithm (NNSA) to estimate the thermodynamic stability of hairpin-loops . This algorithm calculates the Gibbs free energy (ΔG) of the DNA hairpin-loop structure based on known biophysics properties of the base pairs and their interactions[56,57].

The NNSA was applied to the target DNA sequence allowing mismatches in the stem. Each base-pair contributes to the stability of the stem considering the immediate neighboring base-pairs. Adding the local $\Delta G's$ calculated per base and accounting for the size and sequence of the loop results in a final ΔG for the whole hairpin-loop structure. All parameters are available in the literature and the UNAFold web server was used to infer parameters for mismatches[56-58]. For loops larger than 4, no data was available for specific sequences and only the loop size was considered.

Driver hotspot mutations

To identify driver hotspot mutations, all genomic positions with 2 or more mutations were considered for analysis. Hotspot mutations were divided into two groups either located within loops of hairpin-loops or outside of these DNA structures. For hotspot mutations outside of loops, only those in the TpC context were considered, as these are likely initiated by APOBEC enzymes. For hotspot mutations falling within loops, all hotspot mutations in NpC (N = any base) context were considered, as APOBEC3A may also mutate these bases that are not in the TpC context[21].

In cancer, only a few somatic mutations are drivers, while the vast majority are passenger mutations[30]. Under this consideration, we modeled the distribution of the remaining hotspot mutations as a Poisson process per focal hotspot mutation. For more accurate modeling, we considered the tri-nucleotide

context TCW (TCA, TCC, TCG, TCT). In the case of hotspot mutations that do not occur in hairpin loops, DNA accessibility was used as a predictor variable that can influence the distribution of hotspot mutations. We modelled this in R as model_TCW_noloop=glm(n_mut_genpos+DNA_access). Where, n_muts_genpos is a vector with the number of mutations per genomic position linked to the DNA accessibility region (DNA_access). Accessibility regions were divided into 10 regions based on percentiles. Since DNA-accessibility varies per tumor-tissue of origin[59,60], this was estimated for mUC, mBC and other tumor types using ChIPseq experiments from normal tissue as described above.

In the case of hotspot mutations within hairpin-loops, the model was extended to include mutations in non-TCW context (grouped as ApC, CpC or GpC context), the hairpin loop stability (hairpin_stab) and the DNA sequence in the loop (loopSeq): model_TCW_loop=n_mut_genpos+DNA_access+hairpin_stab+loopSeq. The loop sequence was a binary variable to indicate whether the mutation occurred in the following sequence: 1001 or 101 (0 = A/T, 1 = G/C; underlined is the position of the hotspot mutation).

Using these models, we estimated the expected number of mutations of the specified hotspot mutation for which the model was built. Then, the exact Poisson test was applied to estimate the significance of observing the same or more mutations than expected in a specific genomic position. P-values were Benjamini-Hochberg adjusted. In rare occasions, only a few mutations (<2) were available to represent the background distribution of a particular tri-nucleotide. To include these ApoHM in the analysis, a model that did not consider the tri-nucleotide was used instead.

RNA-sequencing

Alignment, pre-processing of RNA-seq data and transcript normalization have been previously described[10,29]. The transcriptomic subtype of each mUC sample was identified when the mean (normalized) expression of all genes associated with a specific subtype was the highest across all subtypes.

mRNA editing

Jalili, et al. identified hotspot mutations in the mRNA of *DDOST* that is targeted by APOBEC3A[17]. The genomic position of this hotspot mutation reveals a hairpin-loop structure that is an ideal substrate for

APOBEC3A. Due to the short life-time of mRNA molecules, the presence of this hotspot mutation reflects ongoing APOBEC mutagenesis. The proportion of C>U mutations in chr1:20981977 was estimated to identify the RNA-editing activity of APOEBC3A.

Transcriptome expression data mapped to genomic regions

MultiBamSummary from deepTools v1.30.0[61] was used to read BAM files and estimate the number of reads in genomic regions with a size of one Mbp. The average raw read count per Mbp was calculated, and a moving average with k = 3 bins was applied. The scale of the read counts was normalized per chromosome using the mean and standard deviation. High transcriptional regions were defined as such when the expression value of one region was above the median of the whole genome.

Simulations and power calculation

A synthetic genome with 1,000,000 hotspot mutations was reconstructed from the original cohort of mUC. To reach the number of hotspot mutations, non-hotspot mutations were randomly selected and the number of mutations per genomic position was drawn from a Poisson distribution using the empirical lambda from the mUC cohort. The same number of driver ApoHM identified in mUC were simulated as hypothetical drivers to replicate a 3%-15% prevalence. Hypothetical cohorts with 10-500 samples were simulated 100 times, using a random number of ApoHM derived from the empirical distribution of the mUC cohort. The statistical power was estimated as the proportion of driver ApoHM that were correctly identified. The performance of the model on simulated cohorts of 100 samples, was also tested with other genomic covariates. These covariates were replication timing and methylation from HeLa cell lines, and the proportion of GC content from the reference build hg19.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

Analyses were performed using the statistical analysis platform R v4.1.0[62]. Fisher's exact, Wilcoxon-rank sum and Wilcoxon signed-rank tests were used for comparison between groups. The correlation coefficients of continuous values with categorical values were estimated with logistic regression analysis applying the Wald test. Residuals for QQ plots and the Kolmogorov–Smirnov test were estimated using

690 DHARMa v.0.4.6[63]. The exact Poisson test was applied to identify potential driver hotspot mutations.
691 The Poisson-binomial method was applied for mutually exclusive mutation events using Rediscover
692 v0.3.2[64] and the Fisher's exact test was applied for the significance of co-occurred mutations. In all
693 cases, p-values were adjusted using the Benjamini-Hochberg method.

694 SUPPLEMENTAL INFORMATION TITLES AND LEGENDS

695

696 Supplementary Figures. Figures S1-S12

697 Supplementary Tables. Tables S1-S7, Related to Figures 3-7

698

References

1. 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. 10.1038/ng.2702.
2. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic. Acids. Res. *47*, D941-D947. 10.1093/nar/gky1015.
3. Boichard, A., Pham, T.V., Yeerna, H., Goodman, A., Tamayo, P., Lippman, S., Frampton, G.M., Tsigelny, I.F., and Kurzrock, R. (2019). APOBEC-related mutagenesis and neo-peptide hydrophobicity: implications for response to immunotherapy. OncoImmunology *8*, 1550341. 10.1080/2162402X.2018.1550341.
4. Rijnders, M., Nakauma-González, J.A., Robbrecht, D.G.J., Gil-Jimenez, A., Balcioglu, H.E., Oostvogels, A.A.M., Aarts, M.J.B., Boormans, J.L., Hamberg, P., van der Heijden, M.S., et al. (2024). Gene-expression-based T-Cell-to-Stroma Enrichment (TSE) score predicts response to immune checkpoint inhibitors in urothelial cancer. Nat. Commun. *15*, 1349. 10.1038/s41467-024-45714-0.
5. Law, E.K., Levin-Klein, R., Jarvis, M.C., Kim, H., Argyris, P.P., Carpenter, M.A., Starrett, G.J., Temiz, N.A., Larson, L.K., Durfee, C., et al. (2020). APOBEC3A catalyzes mutation and drives carcinogenesis in vivo. J. Exp. Med. *217*. 10.1084/jem.20200261.
6. Sharma, S., and Baysal, B.E. (2017). Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. PeerJ *5*, e4136. 10.7717/peerj.4136.
7. Kumar, S., Warrell, J., Li, S., McGillivray, P.D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C.W.Y., Nielsen, M.M., et al. (2020). Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell *180*, 915-927.e916. https://doi.org/10.1016/j.cell.2020.01.032.
8. Shi, M.J., Meng, X.Y., Fontugne, J., Chen, C.L., Radvanyi, F., and Bernard-Pierrot, I. (2020). Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. Genome Med. *12*, 85-85. 10.1186/s13073-020-00781-y.
9. Wong, J.K.L., Aichmüller, C., Schulze, M., Hlevnjak, M., Elgaafary, S., Lichter, P., and Zapatka, M. (2022). Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. Nat. Commun. *13*, 178. 10.1038/s41467-021-27792-6.
10. Nakauma-González, J.A., Rijnders, M., van Riet, J., van der Heijden, M.S., Voortman, J., Cuppen, E., Mehra, N., van Wilpe, S., Oosting, S.F., Rijstenberg, L.L., et al. (2022). Comprehensive Molecular Characterization Reveals Genomic and Transcriptomic Subtypes of Metastatic Urothelial Carcinoma. Eur. Urol. *81*, 331-336. https://doi.org/10.1016/j.eururo.2022.01.026.
11. Chan, K., Roberts, S.A., Klimczak, L.J., Sterling, J.F., Saini, N., Malc, E.P., Kim, J., Kwiatkowski, D.J., Fargo, D.C., Mieczkowski, P.A., et al. (2015). An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat. Genet. *47*, 1067-1072. 10.1038/ng.3378.
12. Roelofs, P.A., Timmermans, M.A.M., Stefanovska, B., den Boestert, M.A., van den Borne, A.W.M., Balcioglu, H.E., Trapman, A.M., Harris, R.S., Martens, J.W.M., and Span, P.N. (2023). Aberrant APOBEC3B Expression in Breast Cancer Is Linked to Proliferation and Cell Cycle Phase. Cells *12*, 1185.
13. Hirabayashi, S., Shirakawa, K., Horisawa, Y., Matsumoto, T., Matsui, H., Yamazaki, H., Sarca, A.D., Kazuma, Y., Nomura, R., Konishi, Y., et al. (2021). APOBEC3B is preferentially expressed at the G2/M phase of cell cycle. Biochem. Biophys. Res. Commun. *546*, 178-184. https://doi.org/10.1016/j.bbrc.2021.02.008.
14. Cortez, L.M., Brown, A.L., Dennis, M.A., Collins, C.D., Brown, A.J., Mitchell, D., Mertz, T.M., and Roberts, S.A. (2019). APOBEC3A is a prominent cytidine deaminase in breast cancer. PLoS Genet. *15*, e1008545-e1008545. 10.1371/journal.pgen.1008545.
15. Glaser, A.P., Fantini, D., Wang, Y., Yu, Y., Rimar, K.J., Podojil, J.R., Miller, S.D., and Meeks, J.J. (2018). APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. Oncotarget *9*, 4537-4548. 10.18632/oncotarget.23344.
16. Robertson, A.G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A.D., Hinoue, T., Laird, P.W., Hoadley, K.A., Akbani, R., et al. (2017). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell *171*, 540-556.e525. 10.1016/J.CELL.2017.09.007.
17. Jalili, P., Bowen, D., Langenbucher, A., Park, S., Aguirre, K., Corcoran, R.B., Fleischman, A.G., Lawrence, M.S., Zou, L., and Buisson, R. (2020). Quantification of ongoing APOBEC3A activity in tumor cells by monitoring RNA editing at hotspots. Nat. Commun. *11*, 2971. 10.1038/s41467-020-16802-8.
18. Allory, Y., Beukers, W., Sagrera, A., Flández, M., Marqués, M., Márquez, M., Van Der Keur, K.A., Dyrskjot, L., Lurkin, I., Vermeij, M., et al. (2014). Telomerase reverse transcriptase promoter mutations in bladder cancer: High frequency across stages, detection in urine, and lack of association with outcome. Eur. Urol. *65*, 360-366. 10.1016/j.eururo.2013.08.052.
19. Siraj, A.K., Bu, R., Iqbal, K., Parvathareddy, S.K., Siraj, N., Siraj, S., Diaz, M.R.F., Rala, D.R., Benito, A.D., Sabido, M.A., et al. (2020). Telomerase reverse transcriptase promoter mutations in cancers derived from multiple organ sites among middle eastern population. Genomics *112*, 1746-1753. https://doi.org/10.1016/j.ygeno.2019.09.017.

20. Buisson, R., Langenbucher, A., Bowen, D., Kwan, E.E., Benes, C.H., Zou, L., and Lawrence, M.S. (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science *364*, eaaw2872-eaaw2872. 10.1126/science.aaw2872.
21. Langenbucher, A., Bowen, D., Sakhtemani, R., Bournique, E., Wise, J.F., Zou, L., Bhagwat, A.S., Buisson, R., and Lawrence, M.S. (2021). An extended APOBEC3A mutation signature in cancer. Nat. Commun. *12*, 1602. 10.1038/s41467-021-21891-0.
22. 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*, 1-12. 10.1038/s41467-019-08576-5.
23. An, L., Dong, C., Li, J., Chen, J., Yuan, J., Huang, J., Chan, K.M., Yu, C.-h., and Huen, M.S.Y. (2018). RNF169 limits 53BP1 deposition at DSBs to stimulate single-strand annealing repair. Proc. Natl. Acad. Sci. U.S.A. *115*, E8286-E8295. doi:10.1073/pnas.1804823115.
24. Cheng, Y.-C., Chiang, H.-Y., Cheng, S.-J., Chang, H.-W., Li, Y.-J., and Shieh, S.-Y. (2020). Loss of the tumor suppressor BTG3 drives a pro-angiogenic tumor microenvironment through HIF-1 activation. Cell Death Dis. *11*, 1046. 10.1038/s41419-020-03248-5.
25. Fischer, J.-P., Els-Heindl, S., and Beck-Sickinger, A.G. (2020). Adrenomedullin – Current perspective on a peptide hormone with significant therapeutic potential. Peptides *131*, 170347. https://doi.org/10.1016/j.peptides.2020.170347.
26. Levine, A.J., and Brivanlou, A.H. (2006). GDF3 at the crossroads of TGF-beta signaling. Cell Cycle *5*, 1069-1073. 10.4161/cc.5.10.2771.
27. Lo, Y.-H., Romes, E.M., Pillon, M.C., Sobhany, M., and Stanley, R.E. (2017). Structural Analysis Reveals Features of Ribosome Assembly Factor Nsa1/WDR74 Important for Localization and Interaction with Rix7/NVL2. Structure *25*, 762-772.e764. https://doi.org/10.1016/j.str.2017.03.008.
28. Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. 10.1038/s41586-020-1969-6.
29. Angus, L., Smid, M., Wilting, S.M., van Riet, J., Van Hoeck, A., Nguyen, L., Nik-Zainal, S., Steenbruggen, T.G., Tjan-Heijnen, V.C.G., Labots, M., et al. (2019). The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. Nat. Genet. *51*, 1450-1458. 10.1038/s41588-019-0507-7.
30. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041.e1021. 10.1016/j.cell.2017.09.042.
31. Juul, R.I., Nielsen, M.M., Juul, M., Feuerbach, L., and Pedersen, J.S. (2021). The landscape and driver potential of site-specific hotspots across cancer genomes. NPJ Genom. Med. *6*, 33. 10.1038/s41525-021-00197-6.
32. Hoopes, James I.I., Cortez, Luis M.M., Mertz, Tony M.M., Malc, Ewa P.P., Mieczkowski, Piotr A.A., and Roberts, Steven A.A. (2016). APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. Cell Rep. *14*, 1273-1282.
33. Kazanov, M.D., Roberts, S.A., Polak, P., Stamatoyannopoulos, J., Klimczak, L.J., Gordenin, D.A., and Sunyaev, S.R. (2015). APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. Cell Rep. *13*, 1103-1109.
34. 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. 10.1038/s41586-022-04972-y.
35. Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C., Grossmann, S., Dawson, K.J., Ju, Y.S., Iorio, F., Tubio, J.M.C., et al. (2019). Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell *176*, 1282-1294.e1220. 10.1016/j.cell.2019.02.012.
36. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94-101. 10.1038/s41586-020-1943-3.
37. Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. *15*, 585-598. 10.1038/nrg3729.
38. Petljak, M., and Maciejowski, J. (2020). Molecular origins of APOBEC-associated mutations in cancer. DNA Repair *94*, 102905. https://doi.org/10.1016/j.dnarep.2020.102905.
39. Patel, M., Nowsheen, S., Maraboyina, S., and Xia, F. (2020). The role of poly(ADP-ribose) polymerase inhibitors in the treatment of cancer and methods to overcome resistance: a review. Cell Biosci. *10*, 35. 10.1186/s13578-020-00390-7.
40. Jekimovs, C., Bolderson, E., Suraweera, A., Adams, M., O'Byrne, K.J., and Richard, D.J. (2014). Chemotherapeutic Compounds Targeting the DNA Double-Strand Break Repair Pathways: The Good, the Bad, and the Promising. Front. Oncol. *4*. 10.3389/fonc.2014.00086.
41. Gillyard, T., and Davis, J. (2021). Chapter Two - DNA double-strand break repair in cancer: A path to achieving precision medicine. In International Review of Cell and Molecular Biology, U. Weyemi, and L. Galluzzi, eds. (Academic Press), pp. 111-137. https://doi.org/10.1016/bs.ircmb.2021.06.003.
42. Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., Chung, H.C., Kindler, H.L., Lopez-Martin, J.A., Miller, W.H., Jr., et al. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker

analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. *21*, 1353-1365. 10.1016/s1470-2045(20)30445-9.

43. Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M.A., Warren, A.Y., Mahbubani, K.T.A., et al. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. Science *370*, 75-82. 10.1126/science.aba8347.
44. Wang, Y., Robinson, P.S., Coorens, T.H.H., Moore, L., Lee-Six, H., Noorani, A., Sanders, M.A., Jung, H., Katainen, R., Heuschkel, R., et al. (2023). APOBEC mutagenesis is a common process in normal human small intestine. Nat. Genet. *55*, 246-254. 10.1038/s41588-022-01296-5.
45. 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. 10.1093/nar/gkx1081.
46. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.
47. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-329. 10.1038/nature14248.
48. Nakauma-González, J.A. (2023). Whole-genome mapping of APOBEC mutagenesis in metastatic urothelial carcinoma identifies driver hotspot mutations and a novel mutational signature. Zenodo. https://doi.org/10.5281/zenodo.10362579.
49. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210-216. 10.1038/s41586-019-1689-y.
50. van Dessel, L.F., van Riet, J., Smits, M., Zhu, Y., Hamberg, P., van der Heijden, M.S., Bergman, A.M., van Oort, I.M., de Wit, R., Voest, E.E., et al. (2019). The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. Nat. Commun. *10*, 1-13. 10.1038/s41467-019-13084-7.
51. Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. Genome Med. *10*, 33-33. 10.1186/s13073-018-0539-0.
52. Hazelaar, D.M., van Riet, J., Hoogstrate, Y., and van de Werken, H.J.G. (2023). Katdetectr: an R/bioconductor package utilizing unsupervised changepoint analysis for robust kataegis detection. GigaScience *12*, giad081. 10.1093/gigascience/giad081.
53. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404. 10.1038/nature11017.
54. Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., Martincorena, I., Dawson, K.J., Iorio, F., Nik-Zainal, S., Bignell, G.R., et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat. Commun. *5*, 2997-2997. 10.1038/ncomms3997.
55. Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M.C., Papaemmanuil, E., Brewer, D.S., Kallio, H.M.L., Högnäs, G., Annala, M., et al. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353-357. 10.1038/nature14347.
56. SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. U.S.A. *95*, 1460-1465. doi:10.1073/pnas.95.4.1460.
57. John SantaLucia, J., and Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. Annu. Rev. Biophys. Biomol. Struct. *33*, 415-440. 10.1146/annurev.biophys.32.110601.141800.
58. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic. Acids. Res. *31*, 3406-3415. 10.1093/nar/gkg595.
59. Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature *518*, 360-364. 10.1038/nature14221.
60. Kübler, K., Karlić, R., Haradhvala, N.J., Ha, K., Kim, J., Kuzman, M., Jiao, W., Gakkhar, S., Mouw, K.W., Braunstein, L.Z., et al. (2019). Tumor mutational landscape is a record of the pre-malignant state. bioRxiv, 517565. 10.1101/517565.
61. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic. Acids. Res. *44*, W160-W165. 10.1093/nar/gkw257.
62. Team, R.C. (2017). R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. R Foundation for Statistical Computing-R Foundation for Statistical Computing.
63. Hartig, F. (2022). DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. https://florianhartig.github.io/DHARMa/.
64. Ferrer-Bonsoms, J.A., Jareno, L., and Rubio, A. (2021). Rediscover: an R package to identify mutually exclusive mutations. Bioinformatics *38*, 844-845. 10.1093/bioinformatics/btab709.