Article

Mutation Hotspots Found in Bladder Cancer Aid Prediction of Carcinogenic Risk in Normal Urothelium

Sydney R. Grant [1,2], Li Tang [3], Lei Wei [4], Barbara A. Foster [5], Gyorgy Paragh [1,2,*] and Wendy J. Huss [1,2,*]

[1] Department of Dermatology, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[2] Department of Cell Stress Biology, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[3] Department of Cancer Prevention & Control, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[4] Department of Biostatistics & Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[5] Department of Pharmacology & Therapeutics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA
[*] Correspondence: gyorgy.paragh@roswellpark.org (G.P.); wendy.huss@roswellpark.org (W.J.H.)

Abstract: More than 80,000 new cases of bladder cancer are estimated to be diagnosed in 2023. However, the 5-year survival rate for bladder cancer has not changed in decades, highlighting the need for prevention. Numerous cancer-causing mutations are present in the urothelium long before signs of cancer arise. Mutation hotspots in cancer-driving genes were identified in non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC) tumor samples. Mutation burden within the hotspot regions was measured in normal urothelium with a low and high risk of cancer. A significant correlation was found between the mutation burden in normal urothelium and bladder cancer tissue within the hotspot regions. A combination of measured hotspot burden and personal risk factors was used to fit machine learning classification models. The efficacy of each model to differentiate between adjacent benign urothelium from bladder cancer patients and normal urothelium from healthy donors was measured. A random forest model using a combination of personal risk factors and mutations within MIBC hotspots yielded the highest AUC of 0.9286 for the prediction of high- vs. low-risk normal urothelium. Currently, there are no effective biomarkers to assess subclinical field disease and early carcinogenic progression in the bladder. Our findings demonstrate novel differences in mutation hotspots in NMIBC and MIBC and provide the first evidence for mutation hotspots to aid in the assessment of cancer risk in the normal urothelium. Early risk assessment and identification of patients at high risk of bladder cancer before the clinical presentation of the disease can pave the way for targeted personalized preventative therapy.

Keywords: urothelium; mutation hotspot; clonal mutations; carcinogenesis; bladder cancer; machine learning

1. Introduction

In 2023, there will be an estimated 82,290 new cases of bladder cancer, with an estimated 16,0 deaths, in the United States [1]. The most common clinical presentation of bladder cancer at the time of diagnosis is visible hematuria [2]. At diagnosis, cancers are classified as non-muscle-invasive bladder cancer (NMIBC) or muscle-invasive bladder cancer (MIBC). Most bladder cancers are NMIBC and are treated by resection with cystoscopy, with or without intravesical immuno- or chemotherapy. The current treatment regimens often cause a considerable physical, psychological, and economic burden. The costs for 5-year NMIBC surveillance range from $50,000 for low-risk patients to over $350,000 for high-risk patients [3]. More advanced stages of bladder cancer, such as MIBC, often require more aggressive treatments, including cystectomy combined with

radiation or chemotherapy [4,5]. Despite advances in oncology, the 5-year survival rate for bladder cancer has not changed in the last 20 years [1], highlighting the need for new strategies in bladder cancer prevention.

Exposure to carcinogens is the highest risk factor for bladder cancer. The most common carcinogen exposures are tobacco and lifetime exposure to chemical byproducts [2,6]. The incidence of bladder cancer also increases with age [2,7]. Carcinogenesis is initiated long before the first clinical signs of cancer arise and requires the accumulation of multiple cancer-driving mutations [8]. Thus, similar to other carcinogen-exposed organs, clinically normal-appearing urothelium harbors numerous genomic mutations [9,10]. Several studies have found a positive selection of mutations in various genes, including *MLL2*, *KDM6A*, *ARID1A*, and *RBM10* [9,10].

Many studies have highlighted the mutational vulnerability of specific genomic regions within cancer-driving genes [11–13]. The regions more sensitive to mutations are referred to as mutation hotspots. Numerous hotspots related to bladder cancer diagnosis and prognosis have been identified, including hotspots in *TERT* [14–16], *TP53* [17,18], *PIK3CA* [15,17,18], and *FGFR3* [14,15,18]. A study comparing mutation hotspots among various stages of bladder cancer revealed an association of *TP53* mutations in more aggressive disease subtypes [15]. The presence of mutations at multiple hotspots has also been validated in normal urothelium [16,17], again supporting the occurrence of mutations before the manifestation of bladder cancer. Our previous publication shows mutation enrichment in normal epidermis within mutation hotspots of cutaneous squamous cell carcinoma [19]. However, no studies thus far provide a comprehensive comparison between highly mutated genomic regions of bladder cancer and normal urothelium. Although most cancer driver mutations cluster at hotspots, studies showing the utility of bladder cancer hotspot mutation analysis to predict cancer risk are lacking. Our work shows how mutation burden can help predict carcinogenic risk in clinically normal urothelium with no noticeable signs of disease. These findings provide data for the future development of genomic assays for individualized bladder cancer risk assessment and targeted bladder cancer prevention.

2. Results

2.1. Overlap between Highly Mutated Genomic Regions in MIBC and NMIBC

We have developed a software tool, hotSPOT, that locates genomic regions with high mutation frequency [20]. Somatic mutation datasets for IMPACT targeted sequencing of 103 NMIBC samples (MSK Eur Urol 2017) [21], and whole-exome sequencing of 409 MIBC samples (TCGA, 2017) [22] were run through hotSPOT as discovery cohorts to identify highly mutated genomic regions in both cancer sub-types. For unbiased comparison between the NMIBC and MIBC datasets, mutations considered from the all datasets were limited to the exome regions included in the MSK-IMPACT sequencing panel, which includes 5080 exons from 341 genes, covering a total of 922,349 bp [23]. HotSPOT was used to locate and rank genomic segments of 100 bp size containing one or more mutations. For each dataset, 10% of all segments containing the most mutations were ranked. Panels of the top 10% most mutated regions in NMIBC consisted of 105 segments (NMIBC10), and in MIBC consisted of 369 segments (MIBC10). The difference in segment quantity between NMIBC10 and MIBC10 is likely due to the increased number of available MIBC samples. There was no statistical difference in mutation frequency between the NMIBC and MIBC samples (p = 0.105, Wilcoxon Rank Sum Test). Several genomic areas contained large clusters of mutations, and, therefore, were covered by multiple genomic segments. A total of 44 genomic regions overlapped between the NMIBC10 and MIBC10 panels. There were 61 genomic regions unique to NMIBC10 and 322 regions unique to MIBC10. A panel of combined genomic regions from NMIBC10 and MIBC10 was also generated to represent the top 10% mutated regions in both types of bladder cancer (BC10). The obtained BC10 panel contained 427 unique genomic regions (Figure 1, Supplementary Figure S1).
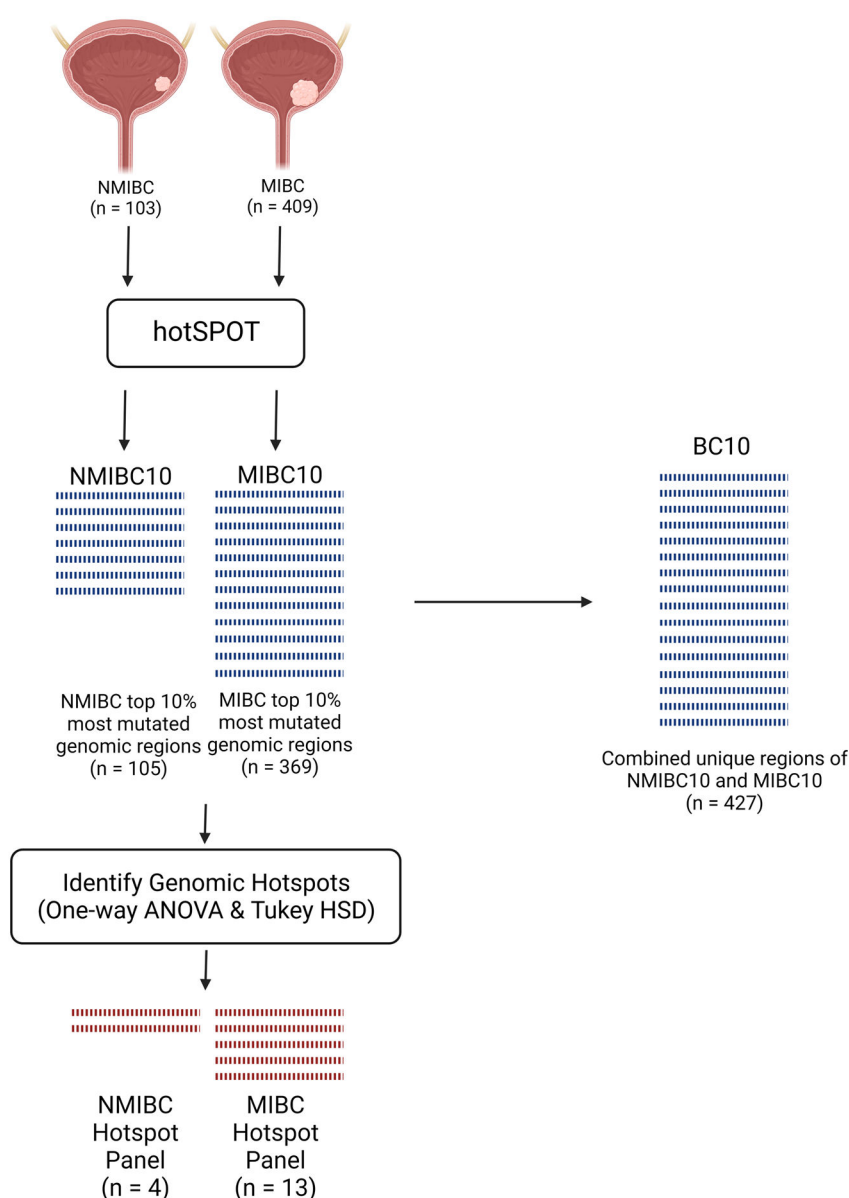
Figure 1. Schematic overview of the establishment of NMIBC and MIBC hotspot panels.

NMIBC10 and MIBC10 panels cover large genomic regions (10,500 bp and 36,900 bp, respectively), making it necessary to develop a refined list of genomic segments representing the most mutated regions of NMIBC and MIBC. Mutation frequency per sample was measured in each genomic segment separately for NMIBC and MIBC. One-way ANOVA was performed for each segment to compare the mutation count in bladder cancer samples amongst all other 100 bp regions within the ranked list. A Tukey HSD test was used to identify all regions that captured significantly more mutations than the majority of lower-ranked regions. This resulted in a final hotspot panel of 4 hotspots in NMIBC and 13 hotspots in MIBC (Figure 1). The 4 hotspots found in the NMIBC dataset were also present in the MIBC hotspot panel.

2.2. Correlation of Mutated Genomic Regions in Bladder Cancer and Normal Urothelium

To better understand the similarities between highly mutated regions in NMIBC and MIBC, we calculated the number of mutations captured in MIBC10 regions in both the NMIBC and MIBC datasets. We found a significant correlation ($p < 0.0001$), indicating that the mutation burden of individual genomic regions as measured with the MIBC10 panel was similar in both datasets. The same analysis using NMIBC10 regions also found a significant correlation ($p < 0.0001$) between the mutation numbers of the hotspot areas in the two different bladder mutations datasets (Figure 2A). These findings demonstrate that mutations are enriched in many common genomic regions in both NMIBC and MIBC.
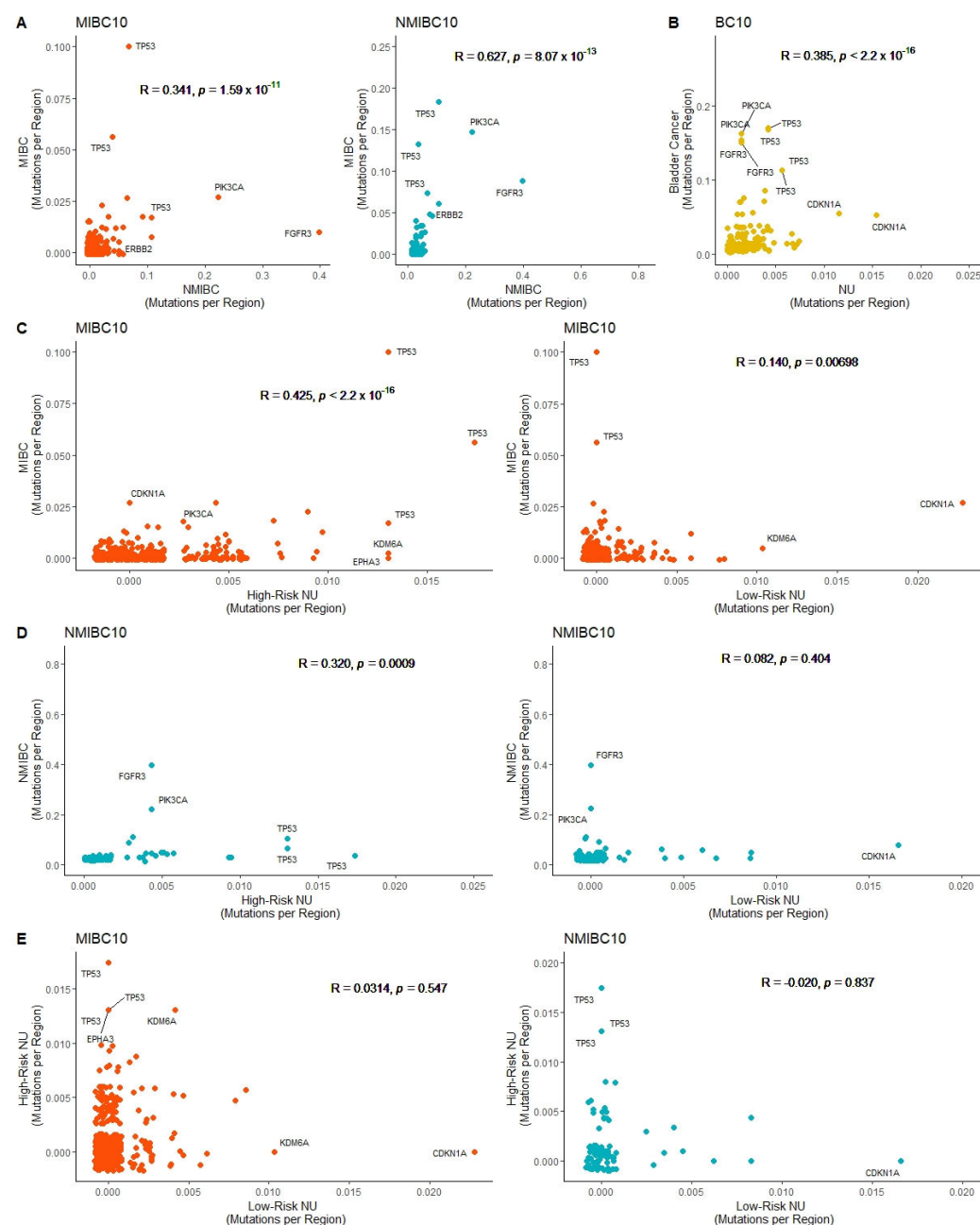

Figure 2. Comparison of mutation burden within bladder cancer hotspots (**A**) Pearson correlation between mutation frequency of NMIBC and MIBC within NMIBC10 and MIBC10 panels. (**B**) Pearson correlation between mutation frequency of NMIBC and MIBC combined and normal urothelium (NU) within BC10 regions. (**C**) Pearson correlation between mutation frequency of MIBC and low-risk NU, MIBC, and high-risk NU within MIBC10 (**D**) Pearson correlation between mutation frequency of NMIBC and low-risk NU, NMIBC, and high-risk NU within NMIBC10. (**E**)

Pearson correlation between mutation frequency of low-risk NU and high-risk NU within MIBC10 and NMIBC10, respectively.

We also assessed whether genomic regions frequently mutated in bladder cancer are also frequently mutated in normal urothelium and how the mutation burden in normal urothelium is associated with an individual's risk of cancer. Previous studies have identified genomic hotspots in bladder cancer [15,17,18] and normal urothelium [14,16]. However, a robust comparison between the genomic hotspots in bladder cancer and normal urothelium is lacking. Furthermore, it is not known if hotspots can be used to identify individuals who are at a high or low risk of developing bladder cancer based on mutation burden in the clinically normal-appearing urothelium. To understand the similarities of highly mutated genomic regions in bladder cancer and clinically normal urothelium, mutation frequency was compared between normal urothelium samples from both healthy donors and bladder cancer patients vs. NMIBC and MIBC samples using BC10. Based on this analysis, a significant correlation ($p < 0.0001$) was found between the mutation frequency in BC10 genomic regions from normal urothelium and bladder cancer samples (Figure 2B).

A previous study that compared normal urothelium from donors with a history of bladder cancer (high risk of bladder cancer) showed an increase in mutations per exome compared with those with no history of bladder cancer (low risk of bladder cancer) [9]. Therefore, we hypothesized that highly mutated regions in high-risk samples showed a higher correlation with mutated regions in bladder cancer. Two publicly available whole-exome sequencing datasets of normal urothelium were identified to compare the mutation rate of hotspots in samples with low- and high-risk bladder cancer [9,10]. The low-risk normal urothelium cohort was comprised of 483 samples from post-mortem donors with no history of cancer from Lawson et al. [9]. The high-risk normal urothelium cohort included normal urothelium collected from individuals with bladder cancer during the time of cystectomy or nephroureterectomy, including 72 samples from Lawson et al. [9] and 161 samples from Li et al. [10].

To compare the mutation landscape between bladder cancer and high- or low-risk normal urothelium, the correlations were calculated between MIBC vs. low-risk normal urothelium and MIBC vs. high-risk normal urothelium within the MIBC10. A significant correlation was found for both comparisons; however, the correlation was more robust in the comparison between high-risk normal urothelium and MIBC samples (R = 0.425 vs. R = 0.140) (Figure 2C). In parallel, we calculated the correlation between NMIBC vs. low-risk normal urothelium and NMIBC vs. high-risk normal urothelium within the NMIBC mutated regions. However, a significant correlation was only found between NMIBC and high-risk normal urothelium ($p < 0.001$) (Figure 2D).

To determine whether the highly mutated regions in high- and low-risk urothelium are similar, the correlation of mutation burden between high- and low-risk normal urothelium within the NMIBC10 and MIBC10 genomic regions was calculated. No significant correlation was found within either of the genomic panels. Based on both panels, hotspots in *KDM6A* and *CDKN1A* were found to be higher in the low-risk normal urothelium, while multiple hotspots of *TP53* were found to be higher in the high-risk normal urothelium (Figure 2E). The apparent differences in mutation distribution of NMIBC10 and MIBC10 regions between high- and low-risk normal urothelium highlight the potential for these genomic regions as markers to differentiate high- and low-risk samples.

2.3. Mutation Hotspots in MIBC and NMIBC

Hotspots included in each hotspot panel were ranked based on the average number of mutations per 100 bp hotspot and compared against the remaining ranked genomic segments in NMIBC10 and MIBC10 (Figure 3A). A ranked list of genes was created based on each hotspot frequency to identify which genes were highly represented in the NMIBC and MIBC hotspot panels. The genes with hotspots in NMIBC were *ERBB2* (1), *FGFR3* (1), *PIK3CA* (1), and *TP53* (1). The genes with the most hotspots in MIBC were *TP53* (5), *FGFR3* (2), *CDKN1A* (1), *CDKN2A* (1), *ERBB2* (1), *KDM6A* (1), *NFE2L2* (1), and *PIK3CA* (Figure 3B). To further validate the relevance of the MIBC hotspot panel, an independent whole-exome sequencing dataset from 126 MIBC samples [10] was used to measure the mutation burden per sample within the hotspot panel and within an equivalent-sized panel of infrequently mutated genomic segments. Based on this analysis, there were significantly more mutations ($p < 0.001$) per sample within the MIBC hotspot panel (Figure 3C). These findings validate using the identified genomic segments as bladder cancer mutation hotspots. The mutation frequency per NMIBC and MIBC sample were measured for each hotspot in both panels. One hotspot in *TP53* was found to have a 3.40-fold increase in average mutation burden in MIBC compared with NMIBC samples ($p < 0.05$) (Figure 3D). However, two hotspots shared between the NMIBC and MIBC hotspot panels were found to have a significantly higher mutation burden in NMIBC samples compared with MIBC. One hotspot in *FGFR3* showed a 4.52-fold increase in average mutations per sample ($p < 0.001$) in NMIBC (Figure 3E), while another in *PIK3CA* showed a 1.52-fold increase in average mutations per sample ($p < 0.05$) in NMIBC (Figure 3F). However, as artificial differences in mutation burden between NMIBC and MIBC samples may be observed due to differences in methods of next-generation sequencing experiments, we tested the validity of the observed mutation differences in the three hotspots. Mutation frequency for each hotspot was measured in samples from the NMIBC samples and an independent dataset of 126 MIBC samples [10]. In the *TP53* hotspot, a 2.03-fold increase in average mutation burden in NMIBC vs. MIBC samples was observed, although this difference did not reach the level of statistical significance in the validation dataset ($p = 0.21$). However, a 4.21-fold increase in average mutations per sample was observed in NMIBC vs. NMIBC samples was shown in the *FGFR3* hotspot ($p < 0.001$) and a 2.36-fold increase in average mutations per sample in NMIBC vs. MIBC samples was shown in the *PIK3CA* hotspot ($p < 0.05$) (Supplementary Figure S3).
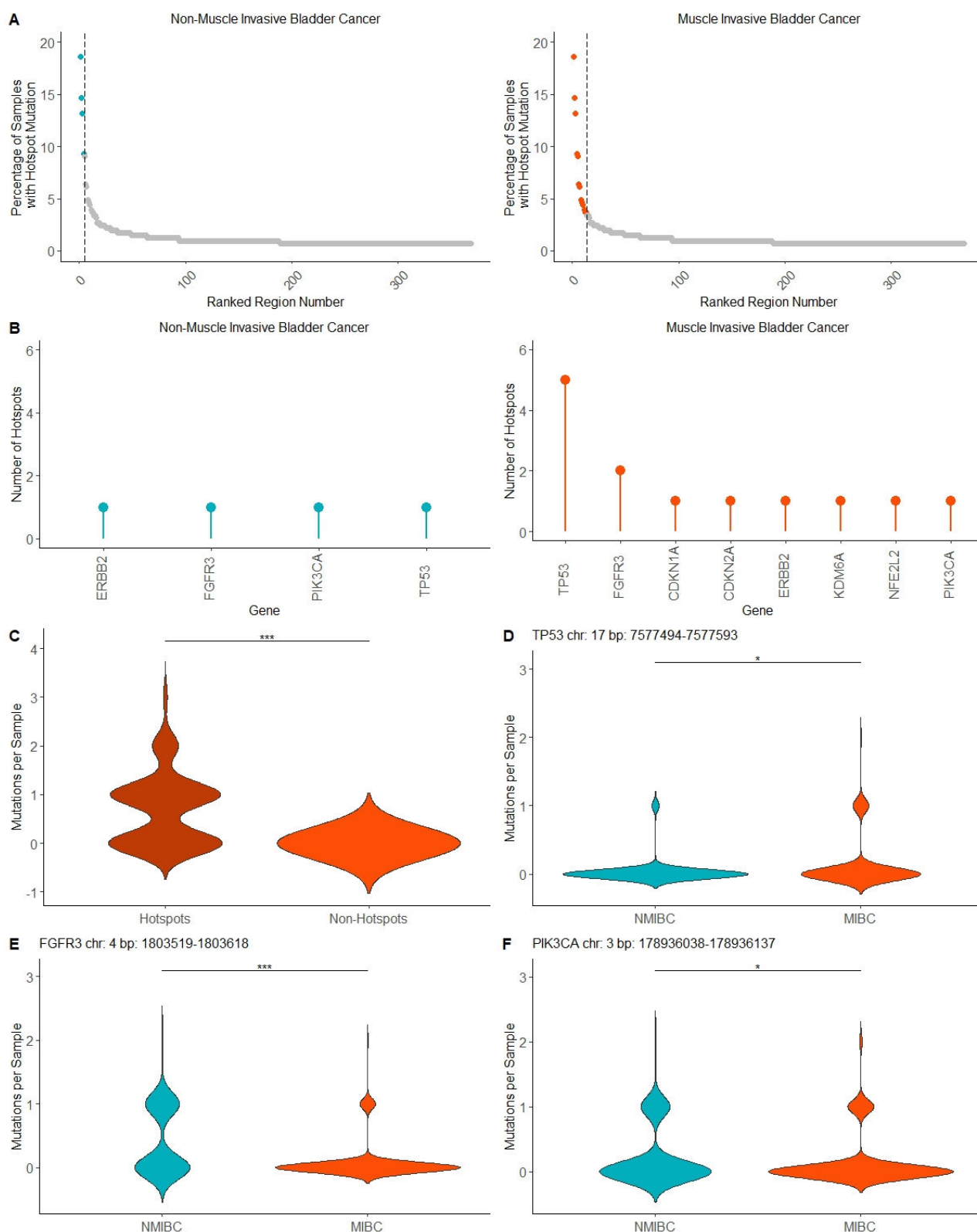
Figure 3. Hotspots in NMIBC and MIBC. (**A**) Percentage of samples with a hotspot mutation for NMIBC and MIBC hotspot panels compared with all genomic regions included in NMIBC10 and MIBC10. (**B**) Frequency of hotspots per gene in NMIBC and MIBC hotspot panels (**C**) Comparison of mutation capture between MIBC hotspots and an equal-sized genomic panel of minimally MIBC mutated regions in external validation MIBC dataset. (**D**) *TP53* hotspot in MIBC hotspot panel with significantly higher mutations in MIBC samples compared with NMIBC samples. (**E**) *FGFR3* hotspot shared between MIBC and NMIBC hotspot panels with significantly higher mutations in NMIBC

samples compared with MIBC samples. (**F**) *PIK3CA* hotspot shared between MIBC and NMIBC hotspot panels with significantly higher mutations in NMIBC samples compared with MIBC samples. Wilcoxon signed-rank test (*: $p < 0.05$, ***: $p < 0.001$).

2.4. Machine Learning Approach to Predicting Bladder Cancer Risk in Normal Urothelium Using Mutation Hotspots

To facilitate using large-scale genomic studies, a machine learning approach to model carcinogenic risk in normal urothelium was used. For the training and testing of our model, the two published datasets of high- and low-risk normal urothelium were used [9,10]. The combined dataset was randomly split 70%/30% into training and test datasets. Each sample in the training dataset was labeled as high or low risk and the mutation burden using the MIBC hotspot panel and NMIBC hotspot panel was calculated. Several personal risk factors for bladder cancer, including age, sex, and smoking status, were also considered in the models. Many published applications of machine learning models have shown that the use of personal risk factors in combination with biological data can improve model performance. Therefore, the personal characteristics known to affect the risk of bladder cancer were included, in addition to mutation hotspot burden, as predictive features for our model. To test the effectiveness of both the personal risk and hotspot features, three different combinations of features were tested: MIBC Hotspot + Personal Risk Factors, NMIBC Hotspots + Personal Risk Factors, and Personal Risk Factors Only. To compare the ability of each feature combination to accurately predict the cancer risk type, each training dataset was fitted to three different types of machine learning classification models. Logistic regression, neural networks, and random forest are the most commonly used machine learning models for assessing genomics and next-generation sequencing data [24,25]. To increase the robustness of our study, we utilized all three model types and compared model performance. The test dataset for each combination of features was then used to assess the model's ability to classify normal urothelium samples based on carcinogenic risk correctly. Personal risk factors alone achieved the lowest AUC values for all three models, with 0.8786, 0.8429, and 0.8964 for logistic regression, neural network, and random forest, respectively. The combination of mutations in NMIBC hotspots and personal risk factors achieved AUC values of 0.8786, 0.8643, and 0.9000. The combination of mutations in MIBC hotspots and personal risk factors showed the best performance for all three models, yielding AUC values of 0.8929, 0.8643, and 0.9286. For all three combinations of features tested, the random forest model showed the strongest performance (Figure 4A). To compare the impact of each feature, the variable importance was calculated for each random forest model. All combinations of features showed age to have the highest importance, while the status of whether a patient was either a current or non-smoker was shown to have the lowest importance (Figure 4B). These findings suggest that the NMIBC and MIBC hotspot panels are able to aid the prediction of cancer risk of clinically normal urothelium.
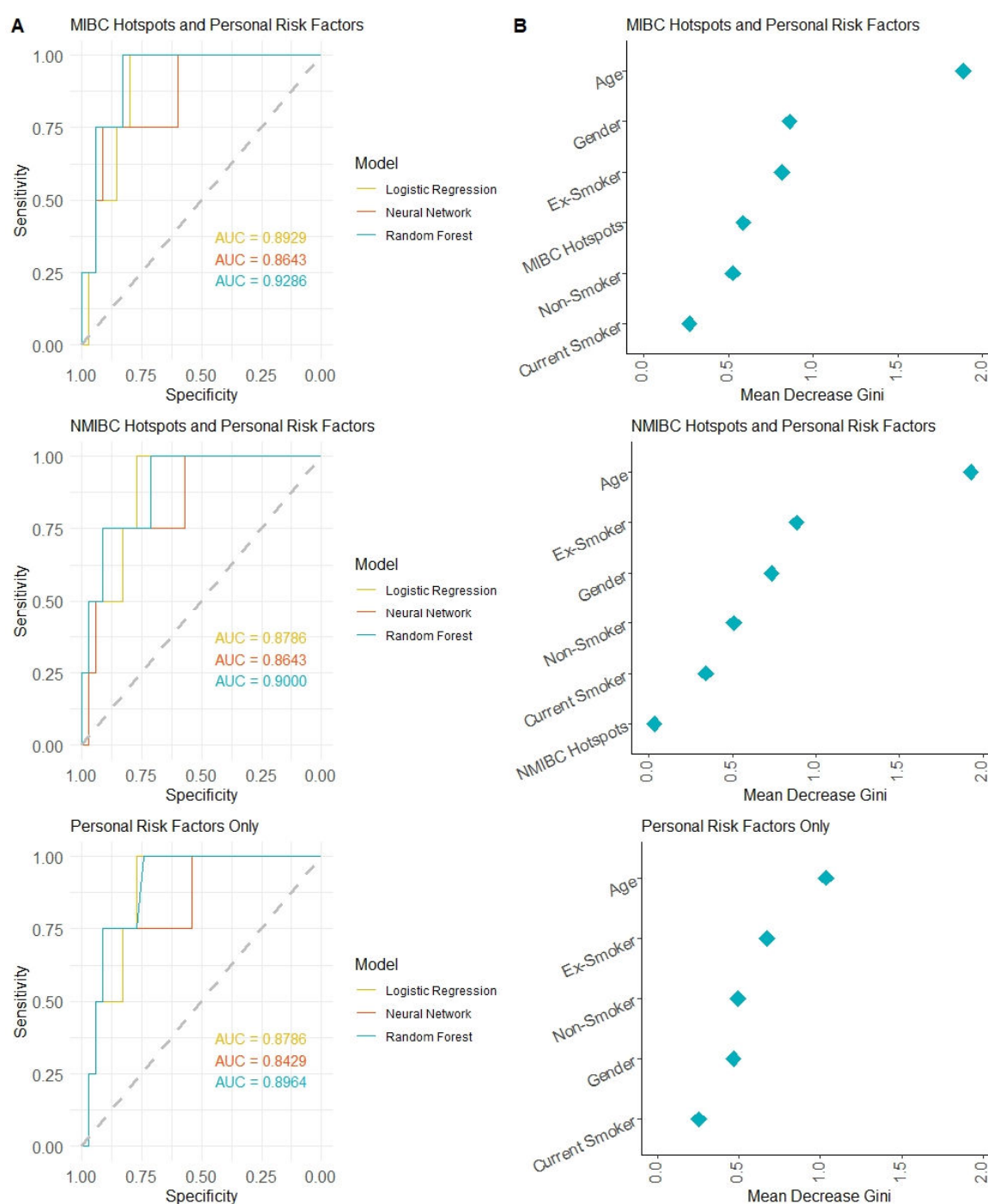
Figure 4. Hotspots in bladder urothelial carcinoma as predictive markers of carcinogenic risk. (**A**) ROC curve of classification model prediction accuracy based on the test dataset for logistic regression, neural network, and random forest models. (**B**) Variable importance plot of the hotspot and clinical variables for cancer risk prediction in the random forest classification model.

3. Discussion

Cancer prevention is moving to the forefront of scientific and clinical interest in the field of oncology. Genomic analysis has the potential to shift our current focus on cancer diagnosis and treatment to a focus on cancer prevention. It is well known that cancer begins to develop long before the first clinical signs of disease [8,26,27]. Despite this knowledge, the early phases of carcinogenesis are understudied. Early identification of individuals at high risk of developing cancer would allow time to implement preventative interventions. Effective interventions may prevent individuals from ever developing

cancer and reduce the treatment burden on those individuals. In this study, mutation hotspots were identified in bladder cancers as valuable tools for assessing carcinogenic risk in the clinically normal-appearing urothelium.

The use of next-generation sequencing has increased in recent years and has become a useful method for studying cancer genomics. However, high-depth sequencing is required to study mutations in clinically normal tissues, which is not currently economically feasible through whole-exome or whole-genome sequencing. Therefore, we have created the hotSPOT tool to identify optimal genomic targets for studying mutations in normal tissues. These studies demonstrate the ability of the hotSPOT tool [20] to identify highly mutated genomic regions among bladder cancer samples. All identified hotspots in NMIBC were also found to be present in MIBC, demonstrating the accumulation of mutation burden during cancer progression.

TP53 was the gene with the largest number of hotspots in MIBC, and a hotspot in *TP53* was also present in the NMIBC hotspot panel. This observation is consistent with previous findings that *TP53* is the most frequently mutated gene with the highest number of mutations in MIBC. *TP53* loss-of-function mutations cause genomic instability and serve as drivers of MIBC [15,28]. While *TP53* loss-of-function mutations are known drivers of aggressive bladder cancer [29], mutations in *TP53* have not previously been associated with early disease or identified as risk factors in the normal urothelium. Somatic mutations in tumor suppressors can be found in all epithelia and can lead to clonal expansion; however, the significance of identifying clonal expansion-associated mutations is still controversial [30]. By middle age, clonal mutations represent half of the normal tissue in the esophagus [31]. In our work, we found not only a higher mutation burden in *TP53* hotspots in MIBC than in NMIBC, but that multiple hotspots of *TP53* were higher in the high-risk normal urothelium. These findings indicate that *TP53* mutations not only play a role in advanced bladder cancer but may serve as a marker of early carcinogenesis.

Hotspots in *CDKN1A*, *CDKN2A*, *KDM6A*, *NFE2L2* were unique to the MIBC hotspot panel. While mutations in these genes are associated with MIBC, recent studies show mutations in normal urothelium and NMIBC. Specifically, *CDKN1A* and *KDM6A* mutations are associated with MIBC [22]; however, recent analysis with a large number of NMIBC samples demonstrates frequent mutations in *CDKN1A* and *KDM6A* in NMIBC [32]. Mutations in *CDKN2A* are more common in MIBC [33], but have recently been identified in normal urothelium and tumor specimens from NMIBC patients [34] and deletion was associated with progression in NMIBC [35]. Mutations in *NFE2L2* (*Nrf*2) are not frequently described as common in bladder cancer; however, analysis of TCGA data of activating mutations found in the *Nrf*2/pathway are common in carcinogen-exposed cancers (including bladder) and are associated with aggressive disease [36]. Therefore, more analysis is needed to identify potential mutations associated with malignancy before tumor initiation.

In addition to *TP53*, hotspot mutations for *ERBB2*, *FGFR3*, and *PIK3CA* were in the NMIBC and MIBC panels. *ERBB2* mutations are less common in NMIBC than MIBC [33]. *FGFR3* mutations are detected in 60–70% of NMIBC cases [37], and *FGFR3* is known for mutation hotspots and translocations [38]. Mutations in *FGFR3* are associated with better prognosis and are considered highly targetable [38,39]; thus, identifying these mutations early could lead to prevention strategies. *PIK3CA* has mutated in approximately 20–30% of NMIBC and MIBC cases. *PIK3CA* mutations are associated with non-smokers, *FGFR3* mutations, and better outcomes in NMIBC [33,40]. Although all NMIBC hotspots were also found in MIBC, NMIBC mutation frequency in hotspots in *FGFR3* and *PIK3CA* was found to be significantly higher than in MIBC. This observation is potentially due to the change in coverage of clonal mutations throughout disease progression. While genes such as *FGFR3* and *PIK3CA* may have a high frequency in the early stages of bladder cancer, and are still detectable, they are outcompeted by cells with driver mutations associated with MIBC.

When assessing the correlation between high- and low-risk normal urothelium in the NMIBC10 and MIBC10, multiple hotspots in both *KDM6A* and *CDKN1A* genes showed higher mutation frequency in low-risk normal urothelium compared with high-risk normal urothelium. The high frequency of hotspot mutations in *KDM6A* and *CDKN1A* in low-risk normal urothelium further indicates that clonal groups of cells with specific mutations can provide initial growth advantage, but can be outcompeted by new, more aggressive clones during disease progression [41]. A further collection of normal urothelium mutation datasets with detailed follow-up of future cancer diagnosis will be necessary to validate the findings in this study and discover additional potential genomic markers of cancer risk.

DNA adducts, particularly 4-aminobiphenyl (cigarette smoke carcinogen) DNA adducts, are higher in the bladder cancers of smokers than in those of non-smokers [42]. A unique aspect of smoking is the introduction of CC > AA dinucleotide to certain cancer types, such as cancers of the lung and esophagus, although the induced mutations are less obvious in bladder cancers because DNA mutation analysis has been performed at low coverage [32]. Neither Lawson et al. nor Li et al. demonstrated a mutation signature associated with smoking [9,10]. Smoking status was not a robust predictive marker of either MIBC or NMIBC. However, this analysis needs further investigation as the Li et al. high-risk dataset did not include any current or former smokers within female patients [10]. However, the common use of aristolochic acid in Chinese patients resulted in a specific mutation pattern of T > A transversions in the 5′-CpTpG-3′ context [10].

Although our findings are compelling, some technical differences between the datasets must be acknowledged. MIBC samples required an average of 60% tumor cell nuclei for samples to pass quality standards, while NMIBC samples required only 40% tumor purity on histologic review. DNA was extracted from fresh frozen MIBC samples and from FFPE in NMIBC samples. Both studies utilized the Illumina HiSeq platform for DNA sequencing. Whole-exome sequencing with 2 × 76 bp paired-end reads was used for MIBC samples, while NMIBC samples were sequenced using the MSK-IMPACT targeted sequencing panel, which includes 2 × 100 bp paired-end reads [23]. Mutation calling was performed using the Firehose and MSK-IMPACT bioinformatics pipelines for MIBC and NMIBC, respectively [21,22]. Due to the variability of next-generation sequencing parameters and samples between the NMIBC and MIBC datasets, further analysis will be needed to validate this study's findings on technically more uniform datasets.

In addition to the technical differences in the analysis of the datasets, the low- and high-risk populations analyzed were different. In Lawson et al., the population was European, and race was not documented [9], compared with the analysis of Chinese patients in Li et al. [10]. As race was not included in the Lawson et al. dataset, we did not include it in our analysis. Future studies could examine how *GSTM1* and *GSTT1* polymorphisms, which alter detoxification of certain carcinogens, alter the mutated genomic regions. *GSTM1*-null mutations affect up to 53% of all racial/ethnic groups, while *GSTT1*-null mutations affect up to 21% of Caucasian, 64% of Asian populations, and 45% of African populations [43,44].

Genomic mutations are known to arise in the earliest phases of carcinogenesis [27,30]. Populations of cells that harbor these mutations are scattered throughout tissues with no clinical signs of disease [8,19,26,45]. The concept of field carcinogenesis is supported by decades of clinical evidence [46]. Despite the long recognition of the mutation burden in normal tissues, our understanding of how these mutations correlate to carcinogenic progression is lacking. Tools to assess early mutations are needed to identify high cancer risk in normal-appearing tissues. Numerous studies have highlighted the efficacy of dietary prevention treatments in bladder cancer [47,48]. However, these methods are widely underutilized. Identifying patients at high risk for developing cancer allows time for the use of preventative interventions to potentially lessen the disease burden on these patients. The development of genomic tools to identify patients who will benefit most

from preventative therapies will help usher in a new era of individualized targeted cancer prevention.

4. Materials and Methods

4.1. Datasets

All datasets used in this study were extracted from publicly available resources (Supplementary Figure S2). The bladder cancer datasets were downloaded from cBioPortal (https://www.cbioportal.org/datasets accessed on 22 November 2022). A single non-muscle invasive bladder cancer dataset was available (MSK Eur Urol 2017) [21]. We chose the muscle-invasive bladder cancer dataset with the largest cohort of whole-exome sequenced samples (TCGA, Cell 2017) [22]. For the discovery of normal urothelium datasets, we searched the terms "bladder", "mutation" and "normal" in PubMed. We screened 405 abstracts published within the last 5 years and identified the two largest datasets of normal urothelium [9,10]. The normal urothelium datasets were divided into two subsets: "high risk" and "low risk". High-risk normal urothelium samples were collected from bladder cancer patients undergoing either cystectomy or nephroureterectomy to remove cancer specimens. Histologically normal areas of tissue were extracted via laser capture microdissection [9,10]. Low-risk normal urothelium samples were microbiopsies collected from post-mortem donors with no history of bladder cancer [9]. The datasets used contained both clinical patient characteristics and single nucleotide variant analysis by DNA-seq. A summary of the datasets, including sample size, sample type, and sequencing parameters, can be found in Supplementary Table S1.

4.2. Development of Hotspot Panels

We have developed an algorithm to design targeted sequencing panels based on genomic areas that harbor high-frequency mutations [20]. This algorithm, hotSPOT, uses SNV data from any sample type and discovers highly mutated areas specific to the inputted dataset. The hotSPOT tool then considers each highly mutated area and identifies an optimal combination of sequencing amplicons to cover these areas. For this study, we developed two distinct hotspot panels for MIBC and NMIBC (Supplementary Tables S2 and S3). Both panels were created using our software tool hotSPOT. The hotSPOT tool is a published RShiny web application and may be accessed at https://rpccc-paraghlab-sgrant.shinyapps.io/hotspot/.

4.3. Analysis in R

All computational analyses were done in R version 4.1.1 [49]. Packages utilized for hotSPOT panel development include hash [50], rlist [51], and R.utils [52]. Packages for data visualization were ggplot2 [53], ggpubr [54], ggsignif [55], and ggrepel [56]. Packages for the development and testing of the cancer risk prediction model were caret [57], glmnet [58], neuralnet [59], randomForest [60], and pROC [61]. All statistical analyses were conducted using R version 4.1.1 'stats' package [49]. A Pearson correlation was calculated to assess the correlation between hotspots. Statistical significance was calculated using a Wilcoxon signed-rank test.

5. Conclusions

The rapidly expanding field of cancer prevention offers a promising new expanding frontier for enhancing patient outcomes. Although numerous studies have reported data on mutation burdens in bladder cancer and normal urothelium, the correlation between mutations and carcinogenic progression remains poorly understood. Prior works comparing high- and low-risk normal urothelium have revealed a higher overall mutation frequency in high-risk samples; however, definitive markers of cancer risk remain elusive. In this study, our computational tool hotSPOT was employed to independently analyze

NMIBC and MIBC datasets. Remarkably, all four hotspots identified in NMIBC were also present in MIBC, despite originating from distinct datasets, and we found the hotspots unique to MIBC as indicators of disease progression. The predictive ability of the MIBC hotspot burden in terms of progression is also supported by the observation that our most effective predictive model incorporated a combination of individual risk factors and mutations within MIBC hotspots. In summary, our findings illuminate novel genomic regions that may play a pivotal role in carcinogenic progression, demonstrating the potential of mutation hotspots to contribute to the development of risk assessment tools for healthy individuals. As our understanding of these genomic landscapes deepens, more effective strategies for cancer prevention and early detection may emerge.

References

1. National Cancer Institute: Surveillance, Epidemiology and End Results Program. Cancer Stat Facts: Bladder Cancer. Available online: https://seer.cancer.gov/statfacts/html/urinb.html (accessed on April 22nd 2023).
2. Lenis, A.T.; Lec, P.M.; Chamie, K.; Mshs, M.D. Bladder Cancer: A Review. *JAMA* **2020**, *324*, 1980–1991. https://doi.org/10.1001/jama.2020.17598.
3. Mossanen, M.; Wang, Y.; Szymaniak, J.; Tan, W.S.; Huynh, M.J.; Preston, M.A.; Trinh, Q.D.; Sonpavde, G.; Kibel, A.S.; Chang, S.L. Evaluating the cost of surveillance for non-muscle-invasive bladder cancer: An analysis based on risk categories. *World J. Urol.* **2019**, *37*, 2059–2065. https://doi.org/10.1007/s00345-018-2550-x.
4. Crabb, S.J.; Douglas, J. The latest treatment options for bladder cancer. *Br. Med. Bull.* **2018**, *128*, 85–95. https://doi.org/10.1093/bmb/ldy034.
5. Patel, V.G.; Oh, W.K.; Galsky, M.D. Treatment of muscle-invasive and advanced bladder cancer in 2020. *CA Cancer J. Clin.* **2020**, *70*, 404–423. https://doi.org/10.3322/caac.21631.
6. Fankhauser, C.D.; Mostafid, H. Prevention of bladder cancer incidence and recurrence: Nutrition and lifestyle. *Curr. Opin. Urol.* **2018**, *28*, 88–92. https://doi.org/10.1097/MOU.0000000000000452.
7. Shariat, S.F.; Sfakianos, J.P.; Droller, M.J.; Karakiewicz, P.I.; Meryn, S.; Bochner, B.H. The effect of age and gender on bladder cancer: A critical review of the literature. *BJU Int.* **2010**, *105*, 300–308. https://doi.org/10.1111/j.1464-410X.2009.09076.x.
8. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med.* **2019**, *11*, 35. https://doi.org/10.1186/s13073-019-0648-4.
9. 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. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **2020**, *370*, 75–82. https://doi.org/10.1126/science.aba8347.

10. Li, R.; Du, Y.; Chen, Z.; Xu, D.; Lin, T.; Jin, S.; Wang, G.; Liu, Z.; Lu, M.; Chen, X.; et al. Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* **2020**, *370*, 82–89. https://doi.org/10.1126/science.aba7300.

11. Chan, K.; Gordenin, D.A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu. Rev. Genet.* **2015**, *49*, 243–267. https://doi.org/10.1146/annurev-genet-112414-054714.

12. Juul, R.I.; Nielsen, M.M.; Juul, M.; Feuerbach, L.; Pedersen, J.S. The landscape and driver potential of site-specific hotspots across cancer genomes. *NPJ Genom. Med.* **2021**, *6*, 33. https://doi.org/10.1038/s41525-021-00197-6.

13. Long, X.; Xue, H. Genetic-variant hotspots and hotspot clusters in the human genome facilitating adaptation while increasing instability. *Hum. Genom.* **2021**, *15*, 19. https://doi.org/10.1186/s40246-021-00318-3.

14. Hayashi, Y.; Fujita, K.; Matsuzaki, K.; Eich, M.L.; Tomiyama, E.; Matsushita, M.; Koh, Y.; Nakano, K.; Wang, C.; Ishizuya, Y.; et al. Clinical Significance of Hotspot Mutation Analysis of Urinary Cell-Free DNA in Urothelial Bladder Cancer. *Front. Oncol.* **2020**, *10*, 755. https://doi.org/10.3389/fonc.2020.00755.

15. Sjodahl, G.; Eriksson, P.; Patschan, O.; Marzouka, N.A.; Jakobsson, L.; Bernardo, C.; Lovgren, K.; Chebil, G.; Zwarthoff, E.; Liedberg, F.; et al. Molecular changes during progression from nonmuscle invasive to advanced urothelial carcinoma. *Int. J. Cancer* **2020**, *146*, 2636–2647. https://doi.org/10.1002/ijc.32737.

16. Weyerer, V.; Eckstein, M.; Strissel, P.L.; Wullweber, A.; Lange, F.; Togel, L.; Geppert, C.I.; Sikic, D.; Taubert, H.; Wach, S.; et al. TERT Promoter Mutation Analysis of Whole-Organ Mapping Bladder Cancers. *Genes* **2021**, *12*, 230. https://doi.org/10.3390/genes12020230.

17. Hovelson, D.H.; Udager, A.M.; McDaniel, A.S.; Grivas, P.; Palmbos, P.; Tamura, S.; Lazo de la Vega, L.; Palapattu, G.; Veeneman, B.; El-Sawy, L.; et al. Targeted DNA and RNA Sequencing of Paired Urothelial and Squamous Bladder Cancers Reveals Discordant Genomic and Transcriptomic Events and Unique Therapeutic Implications. *Eur. Urol.* **2018**, *74*, 741–753. https://doi.org/10.1016/j.eururo.2018.06.047.

18. Shi, M.J.; Meng, X.Y.; Fontugne, J.; Chen, C.L.; Radvanyi, F.; Bernard-Pierrot, I. Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. *Genome Med.* **2020**, *12*, 85. https://doi.org/10.1186/s13073-020-00781-y.

19. Wei, L.; Christensen, S.R.; Fitzgerald, M.E.; Graham, J.; Hutson, N.D.; Zhang, C.; Huang, Z.; Hu, Q.; Zhan, F.; Xie, J.; et al. Ultradeep sequencing differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. *Sci. Adv.* **2021**, *7*, eabd7703. https://doi.org/10.1126/sciadv.abd7703.

20. Grant, S.R.; Rosario, S.R.; Patentreger, A.D.; Shary, N.; Fitzgerald, M.E.; Singh, P.K.; Foster, B.A.; Huss, W.J.; Wei, L.; Paragh, G. HotSPOT: A Computational Tool to Design Targeted Sequencing Panels to Assess Early Photocarcinogenesis. *Cancers* **2023**, *15*, 1612. https://doi.org/10.3390/cancers15051612.

21. Pietzak, E.J.; Bagrodia, A.; Cha, E.K.; Drill, E.N.; Iyer, G.; Isharwal, S.; Ostrovnaya, I.; Baez, P.; Li, Q.; Berger, M.F.; et al. Next-generation Sequencing of Nonmuscle Invasive Bladder Cancer Reveals Potential Biomarkers and Rational Therapeutic Targets. *Eur. Urol.* **2017**, *72*, 952–959. https://doi.org/10.1016/j.eururo.2017.05.032.

22. 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. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **2017**, *171*, 540–556.e525. https://doi.org/10.1016/j.cell.2017.09.007.

23. Cheng, D.T.; Mitchell, T.N.; Zehir, A.; Shah, R.H.; Benayed, R.; Syed, A.; Chandramohan, R.; Liu, Z.Y.; Won, H.H.; Scott, S.N.; et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **2015**, *17*, 251–264. https://doi.org/10.1016/j.jmoldx.2014.12.006.

24. Ghosh, D.; Cabrera, J. Enriched Random Forest for High Dimensional Genomic Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 2817–2828. https://doi.org/10.1109/TCBB.2021.3089417.

25. Tran, K.A.; Kondrashova, O.; Bradley, A.; Williams, E.D.; Pearson, J.V.; Waddell, N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* **2021**, *13*, 152. https://doi.org/10.1186/s13073-021-00968-x.

26. Kennedy, S.R.; Zhang, Y.; Risques, R.A. Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. *Trends Cancer* **2019**, *5*, 531–540. https://doi.org/10.1016/j.trecan.2019.07.007.

27. Risques, R.A.; Kennedy, S.R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet.* **2018**, *14*, e1007108. https://doi.org/10.1371/journal.pgen.1007108.

28. Liu, H.P.; Jia, W.; Kadeerhan, G.; Xue, B.; Guo, W.; Niu, L.; Wang, X.; Wu, X.; Li, H.; Tian, J.; et al. Individualized prognosis stratification in muscle invasive bladder cancer: A pairwise TP53-derived transcriptome signature. *Transl. Oncol.* **2023**, *29*, 101629. https://doi.org/10.1016/j.tranon.2023.101629.

29. Esrig, D.; Elmajian, D.; Groshen, S.; Freeman, J.A.; Stein, J.P.; Chen, S.C.; Nichols, P.W.; Skinner, D.G.; Jones, P.A.; Cote, R.J. Accumulation of nuclear p53 and tumor progression in bladder cancer. *N. Engl. J. Med.* **1994**, *331*, 1259–1264. https://doi.org/10.1056/NEJM199411103311903.

30. Acha-Sagredo, A.; Ganguli, P.; Ciccarelli, F.D. Somatic variation in normal tissues: Friend or foe of cancer early detection? *Ann. Oncol.* **2022**, *33*, 1239–1249. https://doi.org/10.1016/j.annonc.2022.09.156.

31. Martincorena, I.; Fowler, J.C.; Wabik, A.; Lawson, A.R.J.; Abascal, F.; Hall, M.W.J.; Cagan, A.; Murai, K.; Mahbubani, K.; Stratton, M.R.; et al. Somatic mutant clones colonize the human esophagus with age. *Science* **2018**, *362*, 911–917. https://doi.org/10.1126/science.aau3879.

32. Koutros, S.; Rao, N.; Moore, L.E.; Nickerson, M.L.; Lee, D.; Zhu, B.; Pardo, L.A.; Baris, D.; Schwenn, M.; Johnson, A.; et al. Targeted Deep Sequencing of Bladder Tumors Reveals Novel Associations between Cancer Gene Mutations and Mutational Signatures with Major Risk Factors. *Clin. Cancer Res.* **2021**, *27*, 3725–3733. https://doi.org/10.1158/1078-0432.Ccr-20-4419.

33. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **2014**, *507*, 315–322. https://doi.org/10.1038/nature12965.

34. Hayashi, Y.; Fujita, K.; Sakai, K.; Adomi, S.; Banno, E.; Nojima, S.; Tomiyama, E.; Matsushita, M.; Kato, T.; Hatano, K.; et al. Targeted-sequence of normal urothelium and tumor of patients with non-muscle invasive bladder cancer. *Sci. Rep.* **2022**, *12*, 16642. https://doi.org/10.1038/s41598-022-21158-8.

35. Meeks, J.J.; Carneiro, B.A.; Pai, S.G.; Oberlin, D.T.; Rademaker, A.; Fedorchak, K.; Balasubramanian, S.; Elvin, J.; Beaubier, N.; Giles, F.J. Genomic characterization of high-risk non-muscle invasive bladder cancer. *Oncotarget* **2016**, *7*, 75176–75184. https://doi.org/10.18632/oncotarget.12661.

36. Robertson, H.; Dinkova-Kostova, A.T.; Hayes, J.D. NRF2 and the Ambiguous Consequences of Its Activation during Initiation and the Subsequent Stages of Tumourigenesis. *Cancers* **2020**, *12*, 3609. https://doi.org/10.3390/cancers12123609.

37. Hernández, S.; López-Knowles, E.; Lloreta, J.; Kogevinas, M.; Amorós, A.; Tardón, A.; Carrato, A.; Serra, C.; Malats, N.; Real, F.X. Prospective study of FGFR3 mutations as a prognostic factor in nonmuscle invasive urothelial bladder carcinomas. *J. Clin. Oncol.* **2006**, *24*, 3664–3671.

38. Tomlinson, D.; Baldo, O.; Harnden, P.; Knowles, M. FGFR3 protein expression and its relationship to mutation status and prognostic variables in bladder cancer. *J. Pathol.* **2007**, *213*, 91–98. https://doi.org/10.1002/path.2207.

39. Weickhardt, A.J.; Lau, D.K.; Hodgson-Garms, M.; Lavis, A.; Jenkins, L.J.; Vukelic, N.; Ioannidis, P.; Luk, I.Y.; Mariadason, J.M. Dual targeting of FGFR3 and ERBB3 enhances the efficacy of FGFR inhibitors in FGFR3 fusion-driven bladder cancer. *BMC Cancer* **2022**, *22*, 478. https://doi.org/10.1186/s12885-022-09478-4.

40. Dueñas, M.; Martínez-Fernández, M.; García-Escudero, R.; Villacampa, F.; Marqués, M.; Saiz-Ladera, C.; Duarte, J.; Martínez, V.; Gómez, M.J.; Martín, M.L.; et al. PIK3CA gene alterations in bladder cancer are frequent and associate with reduced recurrence in non-muscle invasive tumors. *Mol. Carcinog.* **2015**, *54*, 566–576. https://doi.org/10.1002/mc.22125.

41. Murai, K.; Dentro, S.; Ong, S.H.; Sood, R.; Fernandez-Antoran, D.; Herms, A.; Kostiou, V.; Abnizova, I.; Hall, B.A.; Gerstung, M.; et al. p53 mutation in normal esophagus promotes multiple stages of carcinogenesis but is constrained by clonal competition. *Nat. Commun.* **2022**, *13*, 6206. https://doi.org/10.1038/s41467-022-33945-y.

42. Curigliano, G.; Zhang, Y.J.; Wang, L.Y.; Flamini, G.; Alcini, A.; Ratto, C.; Giustacchini, M.; Alcini, E.; Cittadini, A.; Santella, R.M. Immunohistochemical quantitation of 4-aminobiphenyl-DNA adducts and p53 nuclear overexpression in T1 bladder cancer of smokers and nonsmokers. *Carcinogenesis* **1996**, *17*, 911–916. https://doi.org/10.1093/carcin/17.5.911.

43. Gasper, A.V.; Al-Janobi, A.; Smith, J.A.; Bacon, J.R.; Fortun, P.; Atherton, C.; Taylor, M.A.; Hawkey, C.J.; Barrett, D.A.; Mithen, R.F. Glutathione S-transferase M1 polymorphism and metabolism of sulforaphane from standard and high-glucosinolate broccoli. *Am. J. Clin. Nutr.* **2005**, *82*, 1283–1291. https://doi.org/10.1093/ajcn/82.6.1283.

44. Piacentini, S.; Polimanti, R.; Porreca, F.; Martinez-Labarga, C.; De Stefano, G.F.; Fuciarelli, M. *GSTT1* and *GSTM1* gene polymorphisms in European and African populations. *Mol. Biol. Rep.* **2011**, *38*, 1225–1230. https://doi.org/10.1007/s11033-010-0221-0.

45. Kakiuchi, N.; Ogawa, S. Clonal expansion in non-cancer tissues. *Nat. Rev. Cancer* **2021**, *21*, 239–256. https://doi.org/10.1038/s41568-021-00335-3.

46. Strandgaard, T.; Nordentoft, I.; Lamy, P.; Christensen, E.; Thomsen, M.B.H.; Jensen, J.B.; Dyrskjøt, L. Mutational Analysis of Field Cancerization in Bladder Cancer. *Bladder Cancer* **2020**, *6*, 253–264. https://doi.org/10.3233/BLC-200282.

47. Zhao, H.; Lin, J.; Grossman, H.B.; Hernandez, L.M.; Dinney, C.P.; Wu, X. Dietary isothiocyanates, GSTM1, GSTT1, NAT2 polymorphisms and bladder cancer risk. *Int. J. Cancer* **2007**, *120*, 2208–2213. https://doi.org/10.1002/ijc.22549.

48. Tang, L.; Zirpoli, G.R.; Guru, K.; Moysich, K.B.; Zhang, Y.; Ambrosone, C.B.; McCann, S.E. Intake of cruciferous vegetables modifies bladder cancer survival. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 1806–1811.

49. R Core Team. *R: A Language and Environment for Statistical Computing*, version 4.1.1; R Foundation for Statistical Computing: Vienna, Austria, 2021.

50. Brown, C. *Hash: Full Feature Implementation of Hash/Associated Arrays/Dictionaries*, version 2.2.6.1; R Foundation for Statistical Computing: Vienna, Austria, 2019.

51. Ren, K. *rlist: A Toolbox for Non-Tabular Data Manipulation*, version 0.4.6.2; R Foundation for Statistical Computing: Vienna, Austria, 2021.

52. Bengtsson, H.R. *utils: Various Programming Utilities*, version 2.11.0; R Foundation for Statistical Computing: Vienna, Austria, 2021.

53. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.

54. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*, version 0.4.0; R Foundation for Statistical Computing: Vienna, Austria, 2020.

55. Constantin, A.-E.; Patil, I. *{ggsignif}: R Package for Displaying Significance Brackets for {'ggplot2'}*; R Foundation for Statistical Computing: Vienna, Austria, 2021.

56. Slowikowski, K. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*, version 0.9.1; R Foundation for Statistical Computing: Vienna, Austria, 2021.

57. Kuhn, M. *caret: Classification and Regression Training*, version 6.0-90; R Foundation for Statistical Computing: Vienna, Austria, 2021.
58. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1.
59. Fritsch, S.; Guenther, F.; Wright, M.N. *neuralnet: Training of Neural Networks*, version 1.44.2; R Foundation for Statistical Computing: Vienna, Austria, 2019.
60. Liaw, A.; Wiener, M. *Classification and Regression by randomForest*; Northwestern University: Evanston, IL, USA, 2002.
61. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Muller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77.