Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility

Angeline S. Andrew\textsuperscript{1,}*, Heather H. Nelson\textsuperscript{4}, Karl T. Kelsey\textsuperscript{3}, Jason H. Moore\textsuperscript{1}, Alexis C. Meng\textsuperscript{5}, Daniel P. Casella\textsuperscript{1}, Tor D. Tosteson\textsuperscript{1}, Alan R. Schned\textsuperscript{2} and Margaret R. Karagas\textsuperscript{1}

\textsuperscript{1}Department of Community and Family Medicine, Section of Biostatistics and Epidemiology, \textsuperscript{2}Department of Pathology, \textsuperscript{3}Department of Genetics, Computational Genetics Laboratory, Dartmouth Medical School, Lebanon, NH 03756, USA, \textsuperscript{4}Department of Environmental Health and \textsuperscript{5}Department of Computational Genetics Laboratory, Dartmouth Medical School, Lebanon, NH 03756, USA, \textsuperscript{4}Department of Genetics and Complex Diseases, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA

To whom correspondence should be addressed
Email: Angeline.Andrew@dartmouth.edu

Study results of single nucleotide polymorphisms (SNPs) and cancer susceptibility are often conflicting, possibly because of the analytic challenges of testing for multiple genetic and environmental risk factors using traditional analytic tools. We investigated the relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility in 355 cases and 559 controls enrolled in a population-based study of bladder cancer in the US. Our multifaceted analytical approach included logistic regression, multifactor dimensionality reduction, and hierarchical interaction graphs for the analysis of gene–gene and gene–environment interactions followed by linkage disequilibrium and haplotype analysis. Overall, we did not find an association between any single DNA repair gene SNP and bladder cancer risk. We did find a marginally significant elevated risk of the $XPD$ codon 751 homozygote variant among never smokers [adjusted odds ratio (OR) 2.5, 95% confidence interval (CI) 1.0–6.2]. In addition, the $XRCC1$ 194 variant allele was associated with a reduced bladder cancer risk among heavy smokers [adjusted OR 0.4, 95% CI 0.2–0.9]). The best predictors of bladder cancer included the $XPD$ codon 751 and 312 SNPs along with smoking. Interpretation of this multifactor model revealed that the relationship between the $XPD$ SNPs and bladder cancer is mostly non-additive while the effect of smoking is mostly additive. Since the two $XPD$ SNPs are in significant linkage disequilibrium ($D' = 0.52, P = 0.0001$), we estimated $XPD$ haplotypes. Individuals with variant $XPD$ haplotypes were more susceptible to bladder cancer [e.g. adjusted OR 2.5, 95% CI 1.7–3.6] and the effect was magnified when smoking was considered. These results support the hypothesis that common polymorphisms in DNA repair genes modify bladder cancer risk and emphasize the need for a multifaceted statistical approach to identify gene–gene and gene–environment interactions.

Introduction

Cancer is a multifactorial disease that results from complex interactions between many genetic and environmental factors (1). This is particularly true for the sporadic forms of cancer that, in contrast to familial cancer syndromes, tend to be common in the population. As a result, it is generally believed that there will not be single genes or single environmental factors (i.e. silver bullets) that have large effects on disease susceptibility. Rather, each risk factor is likely to contribute to cancer susceptibility through a combination of nonadditive and additive interactions with other risk factors. This complex genetic architecture is consistent with other common diseases such as cardiovascular disease (2). In fact, it has been suggested that non-additive interactions are a ubiquitous component of the genetic architecture of many common human diseases (3). Given these complexities, a successful research strategy for identifying risk factors for common human cancers must consider combinations of genetic variations and environmental exposures. We describe here a large case–control study of bladder cancer susceptibility that specifically evaluates gene–gene and gene–environment interactions using a multifaceted analytical approach that combines traditional statistical methods with novel computational algorithms.

In 2004, an estimated 60 240 people in the US were diagnosed with bladder cancer and 12 710 died of the disease (4). In the USA, bladder cancer incidence ranks fourth among men, and tenth among women. Occupational exposure to chemicals such as 2-naphthylamine and benzidine, or exposure to 4-aminobiphenyl, and aromatic amines through tobacco smoke, plays a significant role in initiation of bladder cancer. Bladder cancer risk is up to 4-fold higher among cigarette smokers compared with non-smokers (5). Case–control studies provide evidence of a familial predisposition to bladder cancer (6–8) indicating that some susceptibility factors may be heritable.

One such heritable factor is DNA repair polymorphisms that increase susceptibility to DNA damage resulting from bladder carcinogens [reviewed in (9)]. Results of our previous study indicated a 40% reduction in risk of bladder cancer among those with at least one $XRCC1$ 399 variant allele compared with those with one or two wild-type alleles (10). However, to date, epidemiology studies of bladder cancer risk in relation to these polymorphisms are either conflicting (i.e. for $XPD$, $XRCC1$ and $XRCC3$ (11–18), rare (i.e. for $XPC$ PAT) (19), or non-existent (i.e. for $APE1$). Also some studies raise the possibility of gene–gene interaction between polymorphisms, i.e. between $XRCC1$ 194 and $XRCC3$ 241, $XRCC1$ 399/$XRCC1$ 194 and $XPD$ 751/$XPD$ 312 for bladder (18) and lung cancers (20,21). Available data are largely based on hospital-based studies, many of insufficient size to evaluate potential gene–environment interactions. Additionally, differential findings could be related to population admixture. Differences could also be due to the presence of gene–gene and gene–environment interactions that are not well understood due to the analytic

Abbreviations: BER, base excision repair; CI, confidence interval; CVC, cross-validation consistency; DSB, double strand break; MDR, multifactor dimensionality reduction; NER, nucleotide excision repair; OR, odds ratio; SNP, single nucleotide polymorphism.
challenges of testing for multiple genetic and environmental risk factors using traditional analytical tools (22).

To expand on our prior XRCC1 results and clarify the role of polymorphisms in DNA repair genes in bladder cancer susceptibility, we examined multiple single nucleotide polymorphisms (SNPs) in the base excision repair (BER), nucleotide excision repair (NER) and double strand break (DSB) repair pathways in a population-based study of 355 bladder cancer cases and 559 controls from New Hampshire. We specifically evaluated the presence of gene–gene and gene–environment effect modification using both traditional and novel analytic approaches.

Materials and methods

Study group

We identified all cases of bladder cancer diagnosed among New Hampshire residents, ages 25–74 years, from July 1, 1994 to June 30, 1998 from the State Cancer Registry. Within 15 days of diagnosis, the state mandated rapid reporting system requires submission of an initial report of cancer, and a definitive report within 120 days. To be eligible for the study, subjects were required to have a listed telephone number and speak English. We sought physician consent before contacting eligible bladder cancer patients. We interviewed a total of 459 bladder cancer cases, which was 85% of the cases confirmed to be eligible for the study. Non-participants included (n = 10) whose physician denied patient contact, (n = 63) who were reported as deceased by a household member or physician, (n = 3) no answer after 40 attempts distributed over day, evenings and weekends, (n = 75) declined participation and (n = 8) were too ill to take part. A standardized histopathology review was conducted by the study pathologist, and from this review we excluded eleven subjects who were initially reported to the cancer registry as bladder cancer.

All controls were selected from population lists. Controls (n = 665) years of age were selected using population lists obtained from the New Hampshire Department of Transportation. The file contains the names and addresses of those holding a valid driver's license for the state of New Hampshire. Controls 65 years of age and older were chosen from data files provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. The method of control selection used in our study has been successfully employed in other case–control studies conducted in the region [e.g. (23)]. For efficiency, we shared a control group with a study of non-melanoma skin cancer conducted covering an overlapping diagnostic period of July 1, 1993 to June 30, 1995 (23). We selected additional controls for bladder cancer cases diagnosed from July 1, 1995 to June 30, 1997 frequency matched to these cases on age, gender, and from this review we excluded eleven subjects who were initially reported to the cancer registry as bladder cancer.

Personal interview

Informed consent was obtained from each participant and all procedures and study materials were approved by the Committee for the Protection of Human Subjects at Dartmouth College. Consenting participants underwent a detailed in-person interview, usually at their home. Questions covered sociodemographic information (including level of education), lifestyle factors such as use of tobacco (including frequency, duration and intensity of smoking), family history of cancer and medical history prior to the diagnosis date of the bladder cancer cases or reference date assigned to controls. Recruitment procedures for both the shared controls from the non-melanoma skin cancer and additional controls were identical and ongoing concomitantly with the case interviews. Case–control status and the main objectives of the study were not disclosed to the interviewers. To ensure consistent quality of the study interviewers, interviews were tape recorded with the consent of the participants and routinely monitored by the interviewer supervisor. To assess comparability of cases and controls, we asked subjects if they currently held a driver’s license or a Medicare enrollment card.

Genotyping

DNA was isolated from peripheral circulating blood lymphocyte specimens harvested at the time of interview using Qiagen genomic DNA extraction kits (QIAGEN, Valencia, CA). We chose to examine DNA repair genes with polymorphisms that have previously been examined in relation to bladder cancer (XRCC1, XRCC3, XPD, XPC) as well as other pathway members that physically interact with these genes (APE1). Genotyping for non-synonymous SNPs XRCC3 C/T at position 241, APE1 T/G at position 148, XPD G/A at position 1031, XRCC1 C/T at position 194 was performed by Qiagen Genetics using their SNP mass-tagging system. For XRCC1 G/A at position 399 and XPC PAT +/−, genotyping was performed by PCR–RFLP as described previously (10). Of the 1113 participating cases and controls, genotyping was performed on DNA isolated from blood on 914 (82%). For quality control purposes, laboratory personnel were blinded to case–control status. These assays achieved >95% accurate using both negative and positive quality controls (including every 10th sample as a masked duplicate). Data were missing on 103 individuals for XRCC1 194, 70 for XRCC3 399, 2 for XRCC3, 111 for XPD 312, 53 for XPD 751, 131 for XPC and 3 for APE1.

Statistical analysis

The goal of the statistical analysis was to assess the relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. To assess the independent main effects of each SNP, we conducted logistic regression analyses for individuals with one or two variant alleles in comparison to those homozygous wild-type at both loci as the referent group and evaluated interactions between bladder cancer risk factors [gender, smoking variables [e.g. never, <35 pack-years, ≥35 pack-years], and genotype by including interaction terms in a logistic regression model. The pack-year cut-offs were chosen based on the number of published studies. The statistical significances of the interactions were assessed using likelihood ratio tests comparing the models with and without interaction terms.

The nonparametric MDR approach was selected to complement logistic regression for the analysis of gene–gene and gene–environment interactions. We briefly describe MDR here. The details of MDR are described elsewhere (24–27) and reviewed by (28). MDR is a data reduction (i.e. constructive induction) approach that seeks to identify combinations of multilocus genotypes and discrete environmental factors that are associated with either high risk or low risk of disease. Thus, MDR defines a single variable that incorporates information from several loci and/or environmental factors that can be divided into high risk and low risk combinations. This new variable can be evaluated for its ability to classify and predict outcome risk status using cross-validation and permutation testing. With n-fold cross-validation, the data are divided into n equal size pieces. An MDR model is fit using (n − 1)/n of the data (i.e. the training set) and then evaluated for its generalizability on the remaining 1/n of the data (i.e. the testing set). The fitness or value of an MDR model is assessed by estimating accuracy in the training set and the testing set. Accuracy is a function of the percentage of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) and is defined as (TP + TN)/(TP + TN + FP + FN). This process is repeated for all n pieces of the data and the n testing accuracies are averaged to provide an estimate of predictive ability or generalizability. We also estimate the degree to which the same best model is discovered across n divisions of the data. This is referred to as the cross-validation consistency or CV (24,29). A CV of a in n-fold cross-validation is optimal. Here, we selected the best MDR model as the one with the maximum testing accuracy. A testing accuracy of 0.5 is expected under the null hypothesis. Statistical significance is determined using permutation testing. Here, the case–control labels are randomized m times and the entire MDR model fitting procedure repeated on each randomized dataset to determine the expected distribution of testing accuracies under the null hypothesis. It is the combination of cross-validation and permutation testing that reduces the chances of making a type I error due to multiple testing (30,31). In this study, we used 10-fold cross-validation and 1000-fold permutation testing. MDR results were considered statistically significant at the 0.05 level. The MDR software is open-source and freely available from http://www.epistasis.org.

A third approach based on information theory was used to compute the joint interactions and visualize the interactions obtained by logistic regression and MDR. Jakulin and Bratko (32) have provided a metric for determining the gain in information about a class variable (e.g. case–control status) from merging two variables together over that provided by the variables independently (32,33). This measure of information gain allows us to gauge the benefit of considering
two (or more) attributes as one unit. While the concept of information gain is not new (34), its application to the study of variable interactions has been the focus of several recent studies (32,33,35). Consider two variables, A and B, and a class label C. Let \( H(X) \) be the Shannon entropy [see (36)] of \( X \). The information gain (IG) of \( A, B \) and \( C \) can be written as (1) and defined in terms of Shannon entropy (2 and 3).

\[
IG(ABC) = I(A; B|C) - I(A; B) \quad (1)
\]

\[
I(A; B|C) = H(A|C) + H(B|C) - H(A, B|C) \quad (2)
\]

\[
I(A; B) = H(A) + H(B) - H(A, B) \quad (3)
\]

The first term in (1), \( I(IA; BJC) \), measures the interaction of \( A \) and \( B \). The second term, \( I(A; B) \), measures the dependency or correlation between \( A \) and \( B \). If this difference is positive, then there is evidence for an interaction that cannot be linearly decomposed. If the difference is negative, then the information between \( A \) and \( B \) is redundant. If the difference is zero, then there is evidence of conditional independence or a mixture of synergy and redundancy.

These measures of entropy are particularly useful for building interaction graphs that facilitate the interpretation of the relationship between variables. Interaction graphs are comprised of a node for each variable with pairwise connections between them. The percentage of entropy removed (i.e. information gain) by each variable is visualized for each node. The percentage of information gain by each variable is visualized for each node. The percentage of information gain by each variable is visualized for each node. The percentage of information gain by each variable is visualized for each node. The percentage of information gain by each variable is visualized for each node.

**Results**

The study population contained more men than women, and the age distribution was comparable among cases and controls among both sexes (Table I). The majority of the study population was Caucasian (Table I), representing the ethnic make-up of the New Hampshire population. The prevalence of smoking was higher among the cases, as was a first-degree family history of bladder cancer (Table I). The variant allele frequencies for the study population were BER: \( \text{APE1} \) 148 (0.475), \( \text{XRCC1} \) 194 (0.07); DSB: \( \text{XRCC3} \) 241 (0.385); NER: \( \text{XPD} \) 751 (0.37), \( \text{XPD} \) 312 (0.35), \( \text{XPC} \) PAT (0.42).

We began by evaluating the independent effects of each DNA repair SNP on bladder cancer susceptibility using logistic regression. We did not observe that the main effects of the BER polymorphisms at \( \text{APE1} \) 148 and \( \text{XRCC1} \) 194 were related to bladder cancer risk. Among heavy smokers, however, \( \text{XRCC1} \) 194 was associated with a significantly reduced risk of bladder cancer \( [\text{XRCC1} \) 194 heterozygote adjusted OR 0.4, 95% confidence interval (CI) 0.2-0.9]. \( \text{APE1} \) 148 also conferred a slighty reduced risk in the heavy smoking group. Overall the ORs for bladder cancer were not related to smoking status.

---

**Table I. Selected characteristics of bladder cancer cases and controls by gender**

<table>
<thead>
<tr>
<th>Reference age</th>
<th>Men Controls (( n = 360 )) N (%)</th>
<th>Men Cases (( n = 279 )) N (%)</th>
<th>Women Controls (( n = 199 )) N (%)</th>
<th>Women Cases (( n = 76 )) N (%)</th>
<th>Overall Controls (( n = 559 )) N (%)</th>
<th>Overall Cases (( n = 355 )) N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>7 (1.9)</td>
<td>3 (1.1)</td>
<td>19 (9.6)</td>
<td>4 (5.3)</td>
<td>26 (4.7)</td>
<td>7 (2.0)</td>
</tr>
<tr>
<td>40–55</td>
<td>62 (17.2)</td>
<td>36 (12.9)</td>
<td>39 (19.6)</td>
<td>18 (23.7)</td>
<td>101 (18.1)</td>
<td>54 (15.2)</td>
</tr>
<tr>
<td>55–70</td>
<td>198 (55.0)</td>
<td>159 (57.0)</td>
<td>95 (47.7)</td>
<td>32 (42.1)</td>
<td>293 (52.4)</td>
<td>191 (53.8)</td>
</tr>
<tr>
<td>&gt;70</td>
<td>93 (25.8)</td>
<td>81 (29.0)</td>
<td>46 (23.1)</td>
<td>22 (29.0)</td>
<td>139 (24.9)</td>
<td>103 (29.0)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>348 (96.7)</td>
<td>271 (97.1)</td>
<td>195 (98.0)</td>
<td>74 (97.4)</td>
<td>543 (97.1)</td>
<td>345 (97.2)</td>
</tr>
<tr>
<td>Non-white</td>
<td>12 (3.3)</td>
<td>8 (2.9)</td>
<td>4 (2.0)</td>
<td>2 (2.6)</td>
<td>16 (2.9)</td>
<td>10 (2.8)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>173 (48.1)</td>
<td>169 (60.6)</td>
<td>105 (52.8)</td>
<td>44 (57.9)</td>
<td>278 (49.7)</td>
<td>213 (60.0)</td>
</tr>
<tr>
<td>College</td>
<td>187 (51.9)</td>
<td>110 (39.4)</td>
<td>94 (47.2)</td>
<td>32 (42.1)</td>
<td>281 (50.3)</td>
<td>142 (40.0)</td>
</tr>
<tr>
<td>Family history of bladder cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>359 (99.7)</td>
<td>264 (94.6)</td>
<td>194 (97.5)</td>
<td>70 (92.1)</td>
<td>553 (98.9)</td>
<td>334 (94.1)</td>
</tr>
<tr>
<td>Yes</td>
<td>1 (0.3)</td>
<td>15 (5.4)</td>
<td>5 (2.5)</td>
<td>6 (7.9)</td>
<td>6 (1.1)</td>
<td>21 (5.9)</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>85 (23.6)</td>
<td>48 (17.2)</td>
<td>84 (42.2)</td>
<td>22 (29.0)</td>
<td>169 (30.2)</td>
<td>70 (19.7)</td>
</tr>
<tr>
<td>Former</td>
<td>213 (59.2)</td>
<td>150 (53.8)</td>
<td>77 (38.7)</td>
<td>27 (35.5)</td>
<td>290 (51.9)</td>
<td>177 (49.9)</td>
</tr>
<tr>
<td>&lt;35 Pack years*</td>
<td>62 (17.2)</td>
<td>81 (29.0)</td>
<td>38 (19.1)</td>
<td>27 (35.5)</td>
<td>100 (17.9)</td>
<td>108 (30.4)</td>
</tr>
<tr>
<td>≥35 Pack years*</td>
<td>137 (40.2)</td>
<td>79 (28.8)</td>
<td>76 (40.2)</td>
<td>21 (28.8)</td>
<td>213 (40.2)</td>
<td>100 (28.8)</td>
</tr>
</tbody>
</table>

*37 subjects are missing pack-year data.
the DSB polymorphism XRCC3 241, nor was there evidence of a gene–smoking interaction (Table II). In the NER pathway, the XPD 751, XPD 312 or XPC PAT variant genotypes were not associated with an increased bladder cancer risk overall; however, we did observe an elevated risk of the XPD 751 variant among never smokers that was marginally statistically significant [adjusted OR 2.5, 95% CI 1.0–6.2] with a significant gene–smoking interaction (p = 0.04) (Table II). Among individuals who smoked ≥35 pack-years, homozygous variants for XPD 312 had a slightly higher bladder cancer risk [adjusted OR 1.8, 95% CI 0.9–3.9] compared to XPD 312 homozygous wild-types (Table II). XPC PAT, and XPD 751 variant alleles conferred a slightly reduced risk in heavy smokers, however this effect was not statistically significant (Table II).

DNA repair is a complex process involving the cooperation of multiple enzymes in pathways that respond to damage induced by endogenous or exogenous agents, such as tobacco. Therefore, we also evaluated the bladder cancer risk associated with genetic variation in more than one gene and smoking, as described. In the MDR analysis (Table III), pack-years of smoking was the strongest single-factor for predicting bladder cancer risk (average testing accuracy = 0.63, CVC = 10/10). The combination of XPD 751 and XPD 312 was the best two-factor model, with a testing accuracy of 0.65 and a CVC of 8.7/10 (p = 0.001). The three-factor model added pack-years of smoking to XPD 751 and XPD 312 for the most accurate (0.66) model that remained highly consistent in the cross validation (8.6/10) (Table III). All of the four-factor models included XPD 751, XPD 312, and pack-years of smoking. XPC PAT was the most common fourth factor across the 10 datasets, however the addition of this factor decreased the testing accuracy (0.65) and CVC (4.5/10).

After identifying the high risk combinations of factors using MDR, we applied interaction entropy algorithms to facilitate interpretation of the relationship between the variables. As shown in the hierarchical interaction graphs in Figure 1, we found small percentages of the entropy in case–control status explained by the interaction between these two loci (5.74%) (Figure 1A). Pack-years of smoking had a large independent effect (3.78%), however we did not detect a substantial non-additive interaction between smoking and the XPD SNPs considering them individually (Figure 1A), or as a single genotype combination (Figure 1B).

Likewise, the interaction dendrogram (Figure 2) placed XPD 751 and XPD 312 on the same branch. Their position in the diagram indicates that this is the strongest interaction. Pack-years of smoking is located on a different branch than...
the XPD SNPs, supporting the evidence from the interaction entropy graphs that show that there is not a strong relationship between these factors (Figure 2).

We then fit logistic regression models for the independent and joint effects of the XPD polymorphisms in models adjusted for age, gender, and smoking. Compared with individuals who were wild-type at both loci, bladder cancer risk was elevated in individuals who were XPD variant at the 751 locus only [adjusted OR 3.6, 95% CI 2.2–6.3] or XPD variant at the 312 locus only [adjusted OR 5.2, 95% CI 3.0–9.0], but was not as high for variants at both loci (gene–gene interaction \( P < 0.0001 \)). We re-applied interaction entropy algorithms using the XPD 751, XPD 312 genotype combination. The XPD SNPs explained 4.13% of the entropy in case–control (9,11,41,42). We further assessed the association between these factors and other genes in the same or a different pathway. Thus, we suspect that some of the conflicts between the results of previous studies might be due to uncharacterized gene–gene or gene–environment interactions. We addressed this issue by evaluating multiple SNPs in the NER, BER and DSB repair pathways and observed variant allele frequencies that were consistent with those reported in the literature (9,11,41,42). We further assessed the association between genotype, genotype combinations and haplotype with smoking status and bladder cancer risk using multiple traditional and novel statistical approaches.

As more and more studies evaluate risk associated with multiple genes and environmental factors, it has become clear that traditional logistic regression analysis approaches are not adequate for modeling complex multi-factor interactions (43). For this reason, we utilized the recently developed MDR and interaction entropy strategies to assess and interpret potential interactions. This approach improves statistical power to efficiently identify potential gene–gene and gene–environment interactions. The results of these novel algorithms were consistent with our logistic regression analysis for the two-way interaction models. We attempted to test three way interactions to replicate our findings from the MDR analysis in logistic regression; however, the model failed to converge due to the small number of individuals in some cells. Thus, our experience highlights the need for alternative, more powerful methods. Out of all of the possible two-factor combinations tested, MDR analysis selected XPD 751 and XPD 312 as the best two predictors of bladder cancer risk. The three-factor model including XPD 751, XPD 312, and pack-years of

<table>
<thead>
<tr>
<th>Table III. MDR models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best modela</td>
</tr>
<tr>
<td>One factor Pack-years</td>
</tr>
<tr>
<td>Two factors Xpd_751</td>
</tr>
<tr>
<td>Three factors Xpd_312</td>
</tr>
<tr>
<td>Xpd_751</td>
</tr>
<tr>
<td>Xpd_312 Pack-years</td>
</tr>
<tr>
<td>XPD 751 XPD 312</td>
</tr>
<tr>
<td>XPC PAT</td>
</tr>
</tbody>
</table>

aData points for missing values were imputed into 10 independent datasets and these results were consistent for 10/10 imputed datasets.  
Average of most common model over all 10 datasets.

Discussion

The current study demonstrates a comprehensive analytical strategy for investigating risk factors for diseases with a complex genetic architecture. We investigated the hypothesis that prevalent SNPs in DNA repair genes modify genetic susceptibility to bladder cancer. We used a multifaceted analytical approach that combines traditional statistical methods with novel computational algorithms to evaluate gene–gene and gene–environment interactions. The relationship between DNA repair polymorphisms and cancer risk may be particularly complex because the effects of genetic variation in the repair process may depend on the presence of a DNA lesion (e.g. gene–environment interaction) or the presence or absence of polymorphisms in other genes in the same or a different pathway. Thus, we suspect that some of the conflicts between the results of previous studies might be due to uncharacterized gene–gene or gene–environment interactions. We addressed this issue by evaluating multiple SNPs in the NER, BER and DSB repair pathways and observed variant allele frequencies that were consistent with those reported in the literature (9,11,41,42). We further assessed the association between genotype, genotype combinations and haplotype with smoking status and bladder cancer risk using multiple traditional and novel statistical approaches.

As more and more studies evaluate risk associated with multiple genes and environmental factors, it has become clear that traditional logistic regression analysis approaches are not adequate for modeling complex multi-factor interactions (43). For this reason, we utilized the recently developed MDR and interaction entropy strategies to assess and interpret potential interactions. This approach improves statistical power to efficiently identify potential gene–gene and gene–environment interactions. The results of these novel algorithms were consistent with our logistic regression analysis for the two-way interaction models. We attempted to test three way interactions to replicate our findings from the MDR analysis in logistic regression; however, the model failed to converge due to the small number of individuals in some cells. Thus, our experience highlights the need for alternative, more powerful methods. Out of all of the possible two-factor combinations tested, MDR analysis selected XPD 751 and XPD 312 as the best two predictors of bladder cancer risk. The three-factor model including XPD 751, XPD 312, and pack-years of
smoking was the strongest model overall since it had the highest level of testing accuracy and showed good CVC (Table III). Adding other factors (e.g. the four-factor model) lowered the CVC, reduced the testing accuracy and raised the P-value. The MDR three-factor model indicated that XPD, XPD 312, and pack-years of smoking are a high risk combination of factors, but did not specify whether or not there is a synergistic relationship. The interaction entropy and interaction dendogram analyses (Figures 1 and 2) helped us interpret the nature of the interactions in these multifactor models, and revealed that the relationship between the XPD SNPs and bladder cancer is mostly non-additive while the effect of smoking is mostly additive.

XPD is an enzyme in the NER pathway that removes certain DNA crosslinks, UV photolesions, and bulky chemical adducts (44). Non-synonymous SNPs in the XPD gene result in the substitution of glutamine in place of lysine at position 751 and asparagine for aspartic acid at position 312. As reported previously (45), we found that the two SNPs were in linkage disequilibrium, with a higher frequency of XPD Asp, XPD Lys and as such, also examined the risk associated with XPD haplotypes using the PHASE estimation software. Prior, smaller hospital-based studies produced inconsistent results and to date have not examined the bladder cancer risk associated with XPD haplotype (13,15). As in our study, interactions have previously been observed for XPD and 751 in relation to lung cancer risk, and several studies found that the risk of lung cancer associated with the variant allele was higher among non-smokers than among smokers (9,45). In lymphoblastoid cell lines, double variants had an enhanced apoptotic response to UV-induced damage, possibly explaining our findings of an increased risk among those with a variant allele in either XPD 312 or XPD 751 but not for those variant at both loci (20,46). We also observed elevated bladder cancer risk for individuals with the low frequency haplotypes and the bladder cancer ORs for XPD haplotypes also did not vary dramatically by smoking status (Table IV). Thus, future, larger studies of XPD haplotype using more SNPs may be informative.

We also looked at SNPs in other repair pathways, including BER and DSB repair. XRCC1 mediates interactions with...
its BER partners, including APE1, and thereby modulates enzymatic activity throughout the pathway (47). A common amino acid substitution (Arg to Gln) occurs in the BRCT1 domain at codon 399, a region involved in binding polyADPribose polymerase (PARP) and APE1 (47). The relatively uncommon non-synonymous SNP in XRCC1 at position 194 was unrelated to bladder cancer risk overall in our study and others. We observed a lower risk among heavy smokers, similar to what was reported for breast cancer risk (17,48), but with wide confidence intervals. APE1 148, XRCC1 194, XPC PAT and XPD 751 variants also conferred a slightly reduced risk in heavy smokers; however, with the exception of the XRCC1 194 heterozygotes, these effects were not statistically significant.

DSBs may result from replication errors or the action of exogenous agents (9). XRCC3 is required for stabilization of the RAD51 complex in repair of DSBs and cross-links, and for its BER partners, including APE1, and thereby modulates enzymatic activity throughout the pathway (47). A common amino acid substitution (Arg to Gln) occurs in the BRCT1 domain at codon 399, a region involved in binding polyADPribose polymerase (PARP) and APE1 (47). The relatively uncommon non-synonymous SNP in XRCC1 at position 194 was unrelated to bladder cancer risk overall in our study and others. We observed a lower risk among heavy smokers, similar to what was reported for breast cancer risk (17,48), but with wide confidence intervals. APE1 148, XRCC1 194, XPC PAT and XPD 751 variants also conferred a slightly reduced risk in heavy smokers; however, with the exception of the XRCC1 194 heterozygotes, these effects were not statistically significant.

<table>
<thead>
<tr>
<th></th>
<th>Control N (%)</th>
<th>Case N (%)</th>
<th>Adjusted OR (95% CI)</th>
<th>Never smoker</th>
<th>Pack-years &lt;35</th>
<th>Pack-years ≥35</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPD haplotype (312/751)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00 (G/A)</td>
<td>652 (58.3)</td>
<td>369 (52.0)</td>
<td>1.0 Ref</td>
<td>1.0 Ref</td>
<td>1.0 Ref</td>
<td></td>
</tr>
<tr>
<td>01 (G/C)</td>
<td>82 (7.3)</td>
<td>82 (11.6)</td>
<td>1.7 (1.2–2.4)</td>
<td>2.8 (1.4–5.5)</td>
<td>1.8 (1.0–3.2)</td>
<td>1.2 (0.7–2.1)</td>
</tr>
<tr>
<td>10 (A/A)</td>
<td>54 (4.8)</td>
<td>83 (11.7)</td>
<td>2.5 (1.7–3.6)</td>
<td>1.9 (0.8–4.6)</td>
<td>1.4 (0.7–2.7)</td>
<td>4.4 (2.2–8.8)</td>
</tr>
<tr>
<td>11 (A/C)</td>
<td>330 (29.5)</td>
<td>176 (24.8)</td>
<td>0.9 (0.7–1.2)</td>
<td>1.2 (0.8–1.9)</td>
<td>0.7 (0.5–1.1)</td>
<td>0.9 (0.6–1.3)</td>
</tr>
</tbody>
</table>

a Adjusted for age, gender, and smoking (pack-years) (37 subjects are missing pack-year data).
b Adjusted for age and gender.
c OR, odds ratio; CI, confidence interval.
d Ref, reference.

*Refers to single chromosomes.*

Acknowledgements

We would like to thank all members of the New Hampshire Health Study team for making this project possible. This publication was funded in part by National Institute of Health grant numbers CA099500, CA82354, CA57494, from the National Cancer Institute (NCI) and grants ES00002, 5 P42 ES05947, RR018787, and ES07373 from the National Institute of Environmental Health Sciences (NIEHS), NIH. Additional support for Dr. Andrew was kindly provided through a fellowship from the American Society of Preventive Oncology and the Cancer Research Foundation of America. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH, NCI, ASPO, or CRFA.

Conflict of Interest Statement: None declared.