Carcinogenesis vol.27 no.5 pp.1030–1037, 2006
doi:10.1093/carcin/bgi284
Advance Access publication November 25, 2005

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

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

[1]Department of Community and Family Medicine, Section of Biostatistics and Epidemiology, [2]Department of Pathology, [3]Department of Genetics, Computational Genetics Laboratory, Dartmouth Medical School, Lebanon, NH 03756, USA, [4]Department of Environmental Health and [5]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.

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.

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–exposure 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

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 $n = 459$ 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, ($n = 63$) 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 <65 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 (25–34, 35–44, 45–54, 55–64, 65–69, 70–74 years) and gender. Controls were randomly assigned a reference date from among the diagnosis dates of the case group to whom they were matched. We interviewed a total $n = 665$ controls (the total shared control group and additional controls), which was 70% of the controls confirmed to be eligible for the study. Of the potential participants, ($n = 18$) were reported as deceased by a member of the household, ($n = 17$) no answer after 40 attempts distributed over day, evenings and weekends, ($n = 261$) declined, ($n = 29$) were mentally incompetent or too ill to take part.

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 interviewer, 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 enrolment 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 312 and A/C at position 751, XRCC1 C/T at position 194 was performed by Qiagen Genomics 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% accuracy as assessed using and negative and positive quality controls (including every 10th sample as a masked duplicate). Data were missing on 103 individuals for XRCC1 194, 70 for XRCC1 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 for each individual SNP. Assessment of gene–gene and gene–environment interactions was carried out using both logistic regression and Multifactor Dimensionality Reduction (MDR). In addition, we employed a third method that uses information theory to build interaction graphs for confirming, visualizing and interpreting gene–gene and gene–environment interactions identified using logistic regression and MDR. Finally, we carried out linkage disequilibrium and haplotype analyses to assess the allelic effects of predictive SNPs. Each method is described in detail below.

We computed the odds ratio (OR) for the joint effects of gene pairs using individuals who are 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-point was chosen based on the median number of pack-years overall. 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 CVC (24,29). A CVC of n 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 confirm, visualize and interpret the results 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

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

	Men		Women		Overall	
	Controls ($n = 360$) *N* (%)	Cases ($n = 279$) *N* (%)	Controls ($n = 199$) *N* (%)	Cases ($n = 76$) *N* (%)	Controls ($n = 559$) *N* (%)	Cases ($n = 355$) *N* (%)
Reference age						
<40	7 (1.9)	3 (1.1)	19 (9.6)	4 (5.3)	26 (4.7)	7 (2.0)
40–55	62 (17.2)	36 (12.9)	39 (19.6)	18 (23.7)	101 (18.1)	54 (15.2)
55–70	198 (55.0)	159 (57.0)	95 (47.7)	32 (42.1)	293 (52.4)	191 (53.8)
>70	93 (25.8)	81 (29.0)	46 (23.1)	22 (29.0)	139 (24.9)	103 (29.0)
Race						
White	348 (96.7)	271 (97.1)	195 (98.0)	74 (97.4)	543 (97.1)	345 (97.2)
Non-white	12 (3.3)	8 (2.9)	4 (2.0)	2 (2.6)	16 (2.9)	10 (2.8)
Education						
High school or less	173 (48.1)	169 (60.6)	105 (52.8)	44 (57.9)	278 (49.7)	213 (60.0)
≥College	187 (51.9)	110 (39.4)	94 (47.2)	32 (42.1)	281 (50.3)	142 (40.0)
Family history of bladder cancer						
No	359 (99.7)	264 (94.6)	194 (97.5)	70 (92.1)	553 (98.9)	334 (94.1)
Yes	1 (0.3)	15 (5.4)	5 (2.5)	6 (7.9)	6 (1.1)	21 (5.9)
Smoking status						
Never	85 (23.6)	48 (17.2)	84 (42.2)	22 (29.0)	169 (30.2)	70 (19.7)
Former	213 (59.2)	150 (53.8)	77 (38.7)	27 (35.5)	290 (51.9)	177 (49.9)
Current	62 (17.2)	81 (29.0)	38 (19.1)	27 (35.5)	100 (17.9)	108 (30.4)
<35 Pack years[a]	137 (40.2)	79 (28.8)	76 (40.2)	21 (28.8)	213 (40.2)	100 (28.8)
≥35 Pack years[a]	119 (34.9)	147 (53.7)	29 (15.3)	30 (41.1)	148 (27.9)	177 (51.0)

[a]37 subjects are missing pack-year data.

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).

$$\mathrm{IG}(ABC) = I(A; B|C) - I(A; B) \quad (1)$$

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

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

The first term in (1), I(A; B|C), 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 entropy removed for each pairwise Cartesian product of variables is visualized for each connection. Thus, the independent main effects of each SNP, for example, can be quickly compared to the interaction effect. Additive and non-additive interactions can be quickly assessed and used to interpret MDR models that consist of distributions of cases and controls for each genotype combination. A positive entropy (plotted in green) indicates interaction while a negative entropy (plotted in red) indicates redundancy. Interaction entropy analysis was performed using the Orange software package (37). Since the MDR and interaction entropy analysis tools do not permit missing values, missing values were imputed 10 independent times using *S*-plus and analyses were performed using each of the 10 datasets. The results reported were consistent across all 10 datasets.

Interaction dendrograms are also a useful way to visualize interaction (32). Here, hierarchical clustering is used to build a dendrogram that places strongly interacting attributes, as determined by interaction entropy, close together at the leaves of the tree. Jakulin and Bratko (32) define the following dissimilarity measure, $D(A, B)$, that is used by a hierarchical clustering algorithm to build a dendrogram. The value of 1000 is used as an upper bound to scale the dendrograms.

$$D(A, B) = \begin{cases} |I(A; B; C)|^{-1} & \text{if } |I(A; B; C)|^{-1} < 1000 \\ 1000 & \text{otherwise.} \end{cases}$$

Wilke,R.A., Reif,D.M., Moore,J.H. (2005) and Moore,J.H., Gilbert,J.C., Tsai,C.T., Chiang,F.T., Holden,T., Barney,N. and White,B.C. (manuscript submitted) have suggested that this approach will be useful for the analysis and interpretation of gene–gene and gene–environment interactions in genetic and epidemiologic studies (38).

For the two genes with more than one SNP: XPD and XRCC1, we assessed linkage disequilibrium in homozygotes using a chi-square test and inferred haplotypes using PHASE 2.0. This Bayesian method reconstructs the haplotype using Markov chain Monte Carlo techniques by statistically inferring the phase at linked loci from the genotype (39). PHASE is reported to show lower error rates than either the maximum likelihood [expectation maximization algorithm], or the parsimony method (Clark algorithm) (40). Analyses were stratified by age (<50, ≥50), sex and smoking status (ever, never), (never, former, current) or the median smoking intensity (never, <35 pack-years, ≥35 pack- years). We further analyzed the association between genotype and tumor invasiveness (non-invasive versus invasive tumors) using logistic regression, excluding *in-situ* tumors.

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: *APE1* 148 (0.475), *XRCC1* 194 (0.07); DSB: *XRCC3* 241 (0.385); NER: *XPD* 751 (0.37), *XPD* 312 (0.35), *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 *APE1* 148 and *XRCC1* 194 were related to bladder cancer risk. Among heavy smokers, however, *XRCC1* 194 was associated with a significantly reduced risk of bladder cancer [*XRCC1* 194 heterozygote adjusted OR 0.4, 95% confidence interval (CI) 0.2–0.9]. *APE1* 148 also conferred a slightly reduced risk in the heavy smoking group. Overall the ORs for bladder cancer were not related to

Table II. Main effects of genotype on bladder cancer risk overall and by smoking status

	Controls *N* (%)	Cases *N* (%)	Adjusted[a] OR (95% CI)[c]	Never smoker[b] $n = 239$	Pack-years <35[b] $n = 313$	Pack-years ≥35[b] $n = 325$
BER pathway						
APE1 148						
TT	152 (27.3)	101 (28.5)	1.0 Ref[d]	1.0 Ref	1.0 Ref	1.0 Ref
TG	285 (51.2)	186 (52.5)	1.0 (0.7–1.4)	1.3 (0.7–2.5)	1.1 (0.6–1.9)	0.8 (0.5–1.4)
GG	120 (21.5)	67 (18.9)	0.8 (0.5–1.2)	1.0 (0.4–2.4)	1.0 (0.5–2.1)	0.6 (0.3–1.1)
XRCC1 399						
GG	225 (41.8)	118 (38.6)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
GA	227 (42.2)	155 (50.7)	1.4 (1.0–1.9)	1.4 (0.7–2.7)	1.0 (0.6–1.8)	1.6 (1.0–2.6)
AA	86 (16.0)	33 (10.8)	0.8 (0.5–1.2)	0.8 (0.3–2.1)	0.6 (0.3–1.5)	0.8 (0.4–1.8)
GG or GA	452 (84.0)	273 (89.2)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
AA	86 (16.0)	33 (10.8)	0.6 (0.4–1.0)	0.7 (0.3–1.6)	0.6 (0.3–1.4)	0.6 (0.3–1.3)
XRCC1 194						
CC	448 (87.5)	267 (89.3)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
CT	60 (11.7)	29 (9.7)	0.8 (0.5–1.3)	1.7 (0.7–4.1)	1.0 (0.5–2.3)	0.4 (0.2–0.9)
TT	4 (0.78)	3 (1.0)	2.0 (0.4–10.5)	4.8 (0.3–82)	—	0.5 (0.0–5.3)
CT or TT	64 (12.5)	32 (10.7)	0.8 (0.5–1.3)	1.8 (0.8–4.2)	1.1 (0.5–2.4)	0.4 (0.2–0.8)
DSB pathway						
XRCC3 241						
CC	211 (37.9)	146 (41.1)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
CT	272 (48.8)	160 (45.1)	0.9 (0.6–1.2)	0.8 (0.4–1.5)	1.1 (0.7–2.9)	0.7 (0.4–1.1)
TT	74 (13.3)	49 (13.8)	0.9 (0.6–1.4)	0.7 (0.3–1.9)	1.1 (0.5–2.3)	1.0 (0.5–1.9)
NER pathway						
XPC PAT						
−/−	142 (32.6)	132 (37.9)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
+/−	220 (50.6)	158 (45.4)	0.8 (0.6–1.1)	1.0 (0.5–1.9)	0.8 (0.5–1.5)	0.7 (0.4–1.2)
+/+	73 (16.8)	58 (16.7)	0.9 (0.6–1.4)	1.3 (0.5–3.0)	1.1 (0.5–2.3)	0.6 (0.3–1.2)
XPD 312						
GG	205 (39.7)	113 (38.0)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
GA	251 (50.3)	145 (49.1)	1.0 (0.8–1.4)	0.9 (0.5–1.8)	0.7 (0.4–1.3)	1.4 (0.8–2.3)
AA	51 (10.0)	38 (12.9)	1.2 (0.7–2.0)	1.2 (0.4–3.7)	0.7 (0.3–1.7)	1.8 (0.9–3.9)
XPD 751						
AA	210 (38.7)	130 (41.0)	1.0 Ref	1.0 Ref	1.0 Ref	1.0 Ref
AC	268 (49.5)	145 (46.2)	0.9 (0.7–1.2)	1.0 (0.5–1.9)	1.1 (0.6–1.8)	0.8 (0.5–1.2)
CC	66 (11.8)	42 (12.8)	1.0 (0.6–1.6)	2.5 (1.0–6.2)	0.8 (0.3–1.9)	0.6 (0.3–1.3)

[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.

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 *XPD* 751 (0.1%), or *XPD* 312 (0.15%) considered independently, but a large percentage of entropy 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

Table III. MDR models

Best model[a]	Low risk	High risk	Cross-validation consistency[b]	Avg. testing accuracy	*P*-value	Range OR (95% CI)
One factor						
Pack-years	Never smoker or <35 pack-years	≥35 pack-years	10/10	0. 63	<0.002	2.6–2.7 (2.0–3.6)
Two factors						
Xpd_751	Both homozygous wild type, both heterozygous, or both variant	Any other combination	8.7/10	0.65	<0.001	1.7–3.5 (1.5–4.8)
Xpd_312						
Three factors						
Xpd_751	Both homozygous wild type, both heterozygous, or both variant	Any other combination	8.6/10	0.66	<0.001	2.2–4.2 (2.0–5.8)
Xpd_312	Never smoker or <35 pack-years	≥35 pack-years				
Pack-years						
Four factors						
Xpd_751	Any other combination	- Xpd_751 wt, 312 het, XPC wt, smoking	4.5/10	0.65	<0.029	2.1–3.2 (1.8–4.3)
Xpd_312						
Pack-years		- Xpd_751 wt, 312 het, XPC het, ≥35 pack-years				
XPC PAT						

[a]Data points for missing values were imputed into 10 independent datasets and these results were consistent for 10/10 imputed datasets.
[b]Average of most common model over all 10 datasets.

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 status and did not indicate an interaction with pack-years of smoking (Figure 1B). A chi-square test indicated that these two *XPD* loci were in linkage disequilibrium ($P < 0.0001$, $D = 0.12, D' = 0.52$), (while *XRCC1* 399 and *XRCC1* 194 were not). Because of the linkage disequilibrium, we analyzed the *XPD* haplotypes estimated by PHASE in relation to bladder cancer risk using logistic regression with adjustment for age, gender and smoking (shown in Table IV). As in the joint SNP analysis, we found an increased risk for haplotypes with a variant allele at one loci [*XPD* 312 G/751 C, frequency 0.07, adjusted OR 1.7, 95% CI 1.2–2.4; *XPD* 312 A/751 A, frequency 0.05, adjusted OR 2.5, 95% CI 1.7–3.6]. Bladder cancer risk was consistently elevated for individuals with the low frequency haplotypes regardless of smoking status. Among the heavy smokers, bladder cancer risk was associated with a 4-fold bladder cancer risk among those with the *XPD* 312 A/751 A haplotype [adjusted OR 4.4, 95% CI 2.2–8.8].

Risk estimates were not altered significantly when analyses were stratified by level of tumor invasiveness (e.g. invasive versus non-invasive), by type of cancer (e.g. transitional cell versus other), by gender, or restricted to white ethnicity.

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

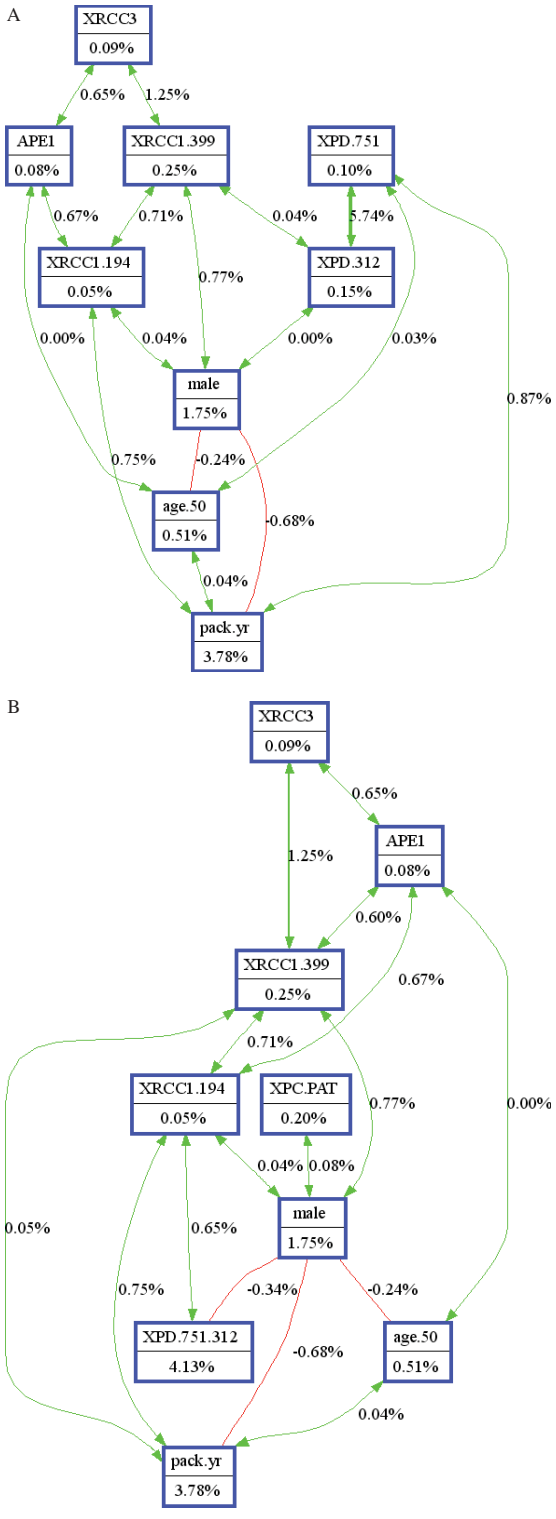

Fig. 1. Orange canvas interaction models. These interaction models describe the percent of the entropy in case–control status that is explained by each factor or two-way interaction. Each gene or environmental factor is shown in a box with the percent of entropy below the label (XPD.751 = *XPD* 751, XPD.312 = *XPD* 312, APE1 = *APE1*, XRCC3 = *XRCC3*, XRCC1.399 = *XRCC1 399*, XRCC1.194 = *XRCC1 194*, male = gender, pack.yr = pack-years of smoking, age.50 = age, XPD.751.312 = *XPD* 751/312 genotype combination). Two-way interactions between factors are depicted as an arrow accompanied by a percent of entropy explained by that interaction. Redundancy is depicted as a line between factors accompanied by a negative percent of entropy. (**A**) The two XPD SNPs (XPD 312 and XPD 751) are included separately in the model, while (**B**) includes the XPD SNPs as a single genotype combination, since they are linked.

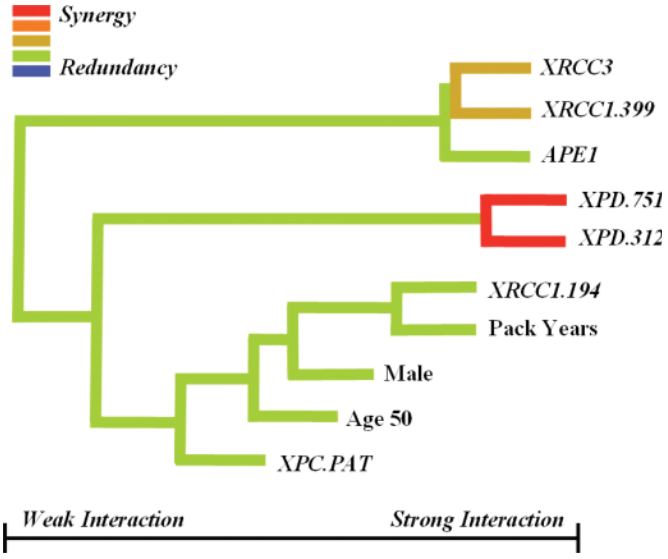

Fig. 2. Interaction dendrogram.

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* 751, *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* 312 Asp, *XPD* 751 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* 312 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

Table IV. Main effects of haplotype on bladder cancer risk overall and by smoking status

	Control *N* (%)[e]	Case *N* (%)[e]	Adjusted[a] OR (95% CI)[c]	Never smoker[b] *N* = 239	Pack-years <35[b] *N* = 313	Pack-years ≥35[b] *N* = 325
XPD haplotype (312/751)						
00 (G/A)	652 (58.3)	369 (52.0)	1.0 Ref[d]	1.0 Ref	1.0 Ref	1.0 Ref
01 (G/C)	82 (7.3)	82 (11.6)	1.7 (1.2–2.4)	2.8 (1.4–5.5)	1.8 (1.0–3.2)	1.2 (0.7–2.1)
10 (A/A)	54 (4.8)	83 (11.7)	2.5 (1.7–3.6)	1.9 (0.8–4.6)	1.4 (0.7–2.7)	4.4 (2.2–8.8)
11 (A/C)	330 (29.5)	176 (24.8)	0.9 (0.7–1.2)	1.2 (0.8–1.9)	0.7 (0.5–1.1)	0.9 (0.6–1.3)

[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.
[e]Refers to single chromosomes.

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 maintaining chromosome stability during cell division (45,49). Results of independent analyses of the DSB pathway *XRCC3* 241 polymorphism and bladder cancer risk have been inconsistent (13–15,18) and we did not observe an increased risk of bladder cancer associated with this polymorphism in our study population, although with limited statistical power. It is possible that *XRCC3* has another biologic function that is modified by the codon 241 amino acid substitution or 241 may be in linkage disequilibrium with another causal polymorphism. Larger studies of multiple SNPs and haplotypes are needed.

Compensatory activity between different DNA repair proteins probably exists (9). Indeed, we observed only a few individuals that were completely wild type for all four of the common DNA repair genes with polymorphisms studied (6 controls and 3 cases). This emphasizes the importance of considering the implications of genetic variation in multiple genes simultaneously. Investigating multiple SNPs also necessitates careful consideration of multiple comparisons issues, a benefit of using the MDR and interaction entropy approaches. Our study involved a number of comparisons, and associations arising out of chance must be considered as a possible explanation for statistically significant results. Further, some of the differences in observed associations across studies may be due to population stratification. The population of New Hampshire is relatively homogeneous and primarily Caucasian, thus, the likelihood of extensive population stratification in our study is generally lower than in more ethnically diverse locations. Restriction to self-reported Caucasian race did not affect our results (data not shown). Future studies employing similar analytic strategies may help to elucidate the impact of population stratification.

In summary, our data indicate that there are potentially important effects of common variations in individual DNA repair genes, particularly in *XPD* on bladder cancer risk and that this risk may be modified by exposure history. We demonstrate the application of a multifaceted analytical approach that produces concordant results and emphasizes the utility of novel bioinformatic analysis tools that make the investigation of gene–gene and gene–environment interactions feasible in a study of reasonable size. Our findings highlight the importance of considering the genetic susceptibility of individuals to complex diseases such as cancer using data on multiple polymorphisms along with a spectrum of potential carcinogenic exposures.

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.

References

1. Pharoah,P.D., Dunning,A.M., Ponder,B.A. and Easton,D.F. (2004) Association studies for finding cancer-susceptibility genetic variants. *Nat. Rev. Cancer*, **4**, 850–860.
2. Sing,C.F., Stengard,J.H. and Kardia,S.L. (2003) Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.*, **23**, 1190–1196.
3. Moore,J.H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, **56**, 73–82.
4. Jemal,A., Tiwari,R.C., Murray,T., Ghafoor,A., Samuels,A., Ward,E., Feuer,E.J. and Thun,M.J. (2004) Cancer statistics, 2004. *CA Cancer J. Clin.*, **54**, 8–29.
5. Silverman,D.T., Morrison,A.S. and Devesa,S.S. (1996) Bladder cancer. In Schottenfeld D. and Fraumeni,J.F. (eds), *Cancer Epidemiology and Prevention*. Oxford University Press, New York, pp. 1156–1179.
6. Cartwright,R.A. (1979) Genetic association with bladder cancer. *Br. Med. J.*, **2**, 798.
7. Sullivan,J.W. (1982) Epidemiologic survey of bladder cancer in greater New Orleans. *J. Urol.*, **128**, 281–283.
8. Kantor,A.F., Hartge,P., Hoover,R.N. and Fraumeni,J.F.Jr. (1985) Familial and environmental interactions in bladder cancer risk. *Int. J. Cancer*, **35**, 703–706.

9. Goode,E.L., Ulrich,C.M. and Potter,J.D. (2002) Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 1513–1530.
10. Kelsey,K.T., Park,S., Nelson,H.H. and Karagas,M.R. (2004) A population-based case–control study of the XRCC1 Arg399Gln polymorphism and susceptibility to bladder cancer. *Cancer Epidemiol. Biomarkers Prev.*, **13**, 1337–1341.
11. Stern,M.C., Johnson,L.R., Bell,D.A. and Taylor,J.A. (2002) XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 1004–1011.
12. Schabath,M.B., Delclos,G.L., Grossman,H.B., Wang,Y., Lerner,S.P., Chamberlain,R.M., Spitz,M.R. and Wu,X. (2005) Polymorphisms in XPD exons 10 and 23 and bladder cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 878–884.
13. Matullo,G., Guarrera,S., Carturan,S., Peluso,M., Malaveille,C., Davico,L., Piazza,A. and Vineis,P. (2001) DNA repair gene polymorphisms, bulky DNA adducts in white blood cells and bladder cancer in a case–control study. *Int. J. Cancer*, **92**, 562–567.
14. Sanyal,S., Festa,F., Sakano,S., Zhang,Z., Steineck,G., Norming,U., Wijkstrom,H., Larsson,P., Kumar,R. and Hemminki,K. (2004) Polymorphisms in DNA repair and metabolic genes in bladder cancer. *Carcinogenesis*, **25**, 729–734.
15. Shen,M., Hung,R.J., Brennan,P., Malaveille,C., Donato,F., Placidi,D., Carta,A., Hautefeuille,A., Boffetta,P. and Porru,S. (2003) Polymorphisms of the DNA repair genes XRCC1, XRCC3, XPD, interaction with environmental exposures, and bladder cancer risk in a case–control study in northern Italy. *Cancer Epidemiol. Biomarkers Prev.*, **12**, 1234–1240.
16. Matullo,G., Palli,D., Peluso,M. *et al*. (2001) XRCC1, XRCC3, XPD gene polymorphisms, smoking and ^{32}P–DNA adducts in a sample of healthy subjects. *Carcinogenesis*, **22**, 1437–1445.
17. Stern,M.C., Umbach,D.M., van Gils,C.H., Lunn,R.M. and Taylor,J.A. (2001) DNA repair gene XRCC1 polymorphisms, smoking, and bladder cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **10**,125–131.
18. Stern,M.C., Umbach,D.M., Lunn,R.M. and Taylor,J.A. (2002) DNA repair gene XRCC3 codon 241 polymorphism, its interaction with smoking and XRCC1 polymorphisms, and bladder cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 939–943.
19. Sak,S.C., Barrett,J.H., Paul,A.B., Bishop,D.T. and Kiltie,A.E. (2005) The polyAT, intronic IVS11-6 and Lys939Gln XPC polymorphisms are not associated with transitional cell carcinoma of the bladder. *Br. J. Cancer*, **92**, 2262–2265.
20. Zhou,W., Liu,G., Miller,D.P., Thurston,S.W., Xu,L.L., Wain,J.C., Lynch,T.J., Su,L. and Christiani,D.C. (2003) Polymorphisms in the DNA repair genes XRCC1 and ERCC2, smoking, and lung cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **12**, 359–365.
21. Chen,S., Tang,D., Xue,K., Xu,L., Ma,G., Hsu,Y. and Cho,S.S. (2002) DNA repair gene XRCC1 and XPD polymorphisms and risk of lung cancer in a Chinese population. *Carcinogenesis*, **23**, 1321–1325.
22. Thornton-Wells,T.A., Moore,J.H. and Haines,J.L. (2004) Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.*, **20**, 640–647.
23. Karagas,M.R., Tosteson,T.D., Blum,J., Morris,J.S., Baron,J.A. and Klaue,B. (1998) Design of an epidemiologic study of drinking water arsenic exposure and skin and bladder cancer risk in a U.S. population. **106** (suppl. 4), 1047–1050.
24. Ritchie,M.D., Hahn,L.W., Roodi,N., Bailey,L.R., Dupont,W.D., Parl,F.F. and Moore,J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.
25. Ritchie,M.D., Hahn,L.W. and Moore,J.H. (2003) Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.*, **24**, 150–157.
26. Hahn,L.W., Ritchie,M.D. and Moore,J.H. (2003) Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, **19**, 376–382.
27. Hahn,L.W. and Moore,J.H. (2004) Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol.*, **4**, 183–194.
28. Moore,J.H. (2004) Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert. Rev. Mol. Diagn.*, **4**, 795–803.
29. Moore,J.H. (2003) Cross-validation consistency for the assessment of genetic programming results in microarray studies. *LNCS*, **2611**, 99–106.
30. Coffey,C.S., Hebert,P.R., Ritchie,M.D., Krumholz,H.M., Gaziano,J.M., Ridker,P.M., Brown,N.J., Vaughan,D.E. and Moore,J.H. (2004) An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene–gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics*, **5**, 49–59.
31. Coffey,C.S., Hebert,P.R., Krumholz,H.M., Morgan,T.M., Williams,S.M. and Moore,J.H. (2004) Reporting of model validation procedures in human studies of genetic interactions. *Nutrition*, **20**, 69–73.
32. Jakulin,A. and Bratko,I. (2003) Analyzing attribute dependencies. In Lavrac,N., Gamberger,D., Blockeel,H. and Todorovski,L. (eds) *PKDD 2003*, *Vol. 2838*, Springer-Verlag, Cavtat, Croatia, pp. 229–240.
33. Jakulin,A., Bratko,I., Smrke,D., Demsar,J. and Zupan,B. (2003) Attribute interactions in medical data analysis. In *Proceedings of the 9th Conference on Artificial Intelligence in Medicine in Europe* (*AIME 2003*), Protaras, Cyprus, October 18–22. Lecture Notes in Artificial Intelligence, Vol. 2780, Springer, pp. 229–238.
34. McGill,W.J. (1954) Multivariate information transmission. *Psychometrika*, **19**, 97–116.
35. Jakulin,A. and Bratko,I. (2004) Testing the significance of attribute interactions. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, July 4–8, 2004. Omnipress, Madison, WI.
36. Pierce,J.R. (1980) *An Introduction to Information Theory—Symbols, Signals and Noise*. Dover Publications, New York.
37. Demsar,J. and Zupan,B. (2004) *Orange: From Experimental Machine Learning to Interactive Data Mining*. White Paper, Faculty of Computer and Information Science, University of Ljubljana, Ljublijana, Slovenia, July 4–8, 2004. July 4–8, 2004. Omnipress, Madison, WI.
38. Wilke,R.A., Reif,D.M. and Moore,J.H. (2005) Combinational pharmacogenetics. *Nat. Rev. Drug Discov.*, **4**, 911–918
39. Stephens,M., Smith,N.J. and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
40. Stephens,M. and Donnelly,P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.
41. Qiao,Y., Spitz,M.R., Guo,Z., Hadeyati,M., Grossman,L., Kraemer,K.H. and Wei,Q. (2002) Rapid assessment of repair of ultraviolet DNA damage with a modified host-cell reactivation assay using a luciferase reporter gene and correlation with polymorphisms of DNA repair genes in normal human lymphocytes. *Mutat. Res.*, **509**, 165–174.
42. Hadi,M.Z., Coleman,M.A., Fidelis,K., Mohrenweiser,H.W. and Wilson,D.M.3rd (2000) Functional characterization of Ape1 variants identified in the human population. *Nucleic Acids Res.*, **28**, 3871–3879.
43. Moore,J.H. and Williams,S.M. (2002) New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.*, **34**, 88–95.
44. Cleaver,J.E. (2000) Common pathways for ultraviolet skin carcinogenesis in the repair and replication defective groups of xeroderma pigmentosum. *J. Dermatol. Sci.*, **23**, 1–11.
45. Butkiewicz,D., Rusin,M., Enewold,L., Shields,P.G., Chorazy,M. and Harris,C.C. (2001) Genetic polymorphisms in DNA repair genes and risk of lung cancer. *Carcinogenesis*, **22**, 593–597.
46. Ronen,A. and Glickman,B.W. (2001) Human DNA repair genes. *Environ. Mol. Mutagen.*, **37**, 241–283.
47. Marsin,S., Vidal,A.E., Sossou,M., Murcia,J.M., Le Page,F., Boiteux,S., De Murcia,G. and Radicella,J.P. (2003) Role of XRCC1 in the coordination and stimulation of oxidative DNA damage repair initiated by the DNA glycosylase hOGG1. *J. Biol. Chem.*, **278**, 44068–44074.
48. Han,J., Hankinson,S.E., De Vivo,I., Spiegelman,D., Tamimi,R.M., Mohrenweiser,H.W., Colditz,G.A. and Hunter,D.J. (2003) A prospective study of XRCC1 haplotypes and their interaction with plasma carotenoids on breast cancer risk. *Cancer Res.*, **63**, 8536–8541.
49. Bishop,D.K., Ear,U., Bhattacharyya,A., Calderone,C., Beckett,M., Weichselbaum,R.R. and Shinohara,A. (1998) Xrcc3 is required for assembly of Rad51 complexes *in vivo*. *J. Biol. Chem.*, **273**, 21482–21488.

Received September 13, 2005; revised November 15, 2005; accepted November 20, 2005