Comprehensive Pathway-Based Association Study of DNA Repair Gene Variants and the Risk of Nasopharyngeal Carcinoma

Hai-De Qin, Yin Yao Shugart, Jin-Xin Bei, Qing-Hua Pan, Lina Chen, Qi-Sheng Feng, Li-Zhen Chen, Wei Huang, Jian Jun Liu, Timothy J. Jorgensen, Yi-Xin Zeng, and Wei-Hua Jia

Abstract

DNA repair plays a central role in protecting against environmental carcinogenesis, and genetic variants of DNA repair genes have been reported to be associated with several human malignancies. To assess whether DNA gene variants were associated with nasopharyngeal carcinoma (NPC) risk, a candidate gene association study was conducted among the Cantonese population within the Guangdong Province, China, the ethnic group with the highest risk for NPC. A 2-stage study design was utilized. In the discovery stage, 676 tagging SNPs covering 88 DNA repair genes were genotyped in a matched case-control study (cases/controls = 755/755). Eleven SNPs with \( P \text{combined} < 0.01 \) were identified. Seven of these SNPs were located within 3 genes, RAD51L1, BRCA2, and TP53BP1. In the validation stage, these 11 SNPs were genotyped in a separate Cantonese population (cases/controls = 1,568/1,297). Two of the SNPs (rs927220 and rs11158728), both in RAD51L1, remained strongly associated with NPC. The SNP rs927220 had a significant \( P \text{combined} \) of 5.55 \( \times 10^{-5} \), with \( OR = 1.20 \) (95% CI = 1.10–1.30). Bonferroni corrected \( P = 0.0381 \). The other SNP (rs11158728), which is in strong linkage disequilibrium with rs927220 (\( r^2 = 0.7 \)), had a significant \( P \text{combined} \) of 2.0 \( \times 10^{-4} \), Bonferroni corrected \( P = 0.1372 \). Gene–environment interaction analysis suggested that the exposures of salted fish consumption and cigarette smoking had potential interactions with DNA repair gene variations, but need to be further investigated. Our findings support the notion that DNA repair genes, in particular RAD51L1, play a role in NPC etiology and development. Cancer Res; 71(8); 1–9. ©2011 AACR.

Introduction

Nasopharyngeal carcinoma (NPC) is rare in most parts of the world, but is a leading malignancy in southern China and Southeast Asia, with the highest incidence rate (40 per 100,000 person-years) among the Cantonese-speaking subpopulation within a region of the Guangdong province, China (1). The distinctive geographic and ethnic distribution of NPC suggests that NPC is a malignancy with complex etiology involving both genetic and environmental factors (2, 3). The risk factors most strongly associated with NPC incidence have been shown to be Epstein-Barr virus (EBV) infection and consumption of salt-preserved fish (4). However, other salt-preserved foods (5) and cigarette smoking (6) have also been consistently reported to be moderate risk factors.

Environmental carcinogens such as N-nitrosamines, polycyclic aromatic hydrocarbons, and aromatic amines, are known components of salt-preserved foods and tobacco (7). These chemicals form DNA adducts, which can result in carcinogenic mutations if left unrepaired. In particular, nitrosamine metabolism-related DNA damage has been recently linked to NPC (8). Furthermore, the metabolites of these carcinogens can generate reactive oxygen species (ROS), which in turn produce base damage, single-strand breaks (SSBs) and double-strand breaks (DSBs) in DNA (9).

There is a growing body of evidence that EBV may promote DNA damage (10–14). These multiple reports of DNA damage promotion by EBV, taken together with the known DNA damaging activity of the NPC-associated chemical carcinogens, support the notion that DNA damage may contribute to NPC induction.
Variants and mutations in various DNA repair genes have been associated with several cancers (15–17). Yet, the prior studies with positive findings for NPC associations have targeted only a limited number of DNA repair genes (18, 19). Because environmental risk factors for NPC contribute to DNA damage, we hypothesized that individuals with inherited deficiencies in DNA repair genes would be more prone to NPC. In this investigation, we sought to determine whether common variation in DNA repair genes is associated with NPC, either alone or in combination with nongenetic NPC risk factors, among individuals living in the Guangdong Province, China, an NPC endemic region.

Materials and Methods

Study subjects

This study included NPC patients who were enrolled at the Sun Yat-Sen University Cancer Center (SYSUCC) between October 2005 and October 2007. Clinical information, including age of onset, histopathologic diagnosis and clinical staging were obtained from medical records. All cases were born and had continuously lived in the Guangdong province for at least 5 years. Cases were histopathologically confirmed for NPC, and were without any prior history of cancer, immunological disease, or mental disease. Once the patients had consented to join the study, interviews were performed and peripheral blood was drawn. In the end, a total of 1,948 eligible NPC cases were identified. Among them, 1,845 interviews were conducted; 103 of them could not complete the questions. The majority of cases (1,767 cases, 95.8%) were interviewed less than 6 months after diagnosis of NPC. A subset of 755 newly diagnosed cases (1,767 cases, 95.8%) were interviewed less than 6 months (average 28 days since diagnosis, ranging from 1 to 106 days). Cantonese-speaking cases were randomly selected from the database to become the case group for the study.

Healthy controls were recruited from 21 health clinics throughout the Guangdong province. In addition, 35 rural villages across the Guangdong province were randomly selected for recruitment to obtain an urban/rural residence distribution similar to cases. Each healthy subject was selected randomly and contacted by telephone. If they consented to participate, they were interviewed in person and blood was drawn. The inclusion/exclusion criteria were the same as for cases. The consent rates were 96% for the cases and 66% for controls. A subset of healthy controls were then randomly selected from the database, matched 1:1 with the selected cases by age (±5 years), sex, geographic region, and dialect.

This study was approved by the Human Ethics Committee, Sun Yat-Sen University. Informed consent was obtained from each participant at recruitment.

Epidemiology data collection

During 30-min interviews, trained staff interviewers collected data on demographics, dietary habits, cigarette smoking, family history of cancer, and other potential cancer risk factors. Regarding salted fish intake, subjects were asked to choose from 5 intake frequency categories (never, sometimes, monthly, weekly, and daily). The reference point for childhood intake was 6 to 12 years old; the reference point for adulthood intake was either the 3 years before diagnosis (cases) or the 3 years before interview (controls). For smoking status, individuals who had smoked at least 100 cigarettes in their lifetime were defined as “smokers.” Smokers with cumulative cigarette smoking exposure of <20 pack-years were defined as “light smokers,” while ≥20 pack-years were “heavy smokers.”

DNA was extracted from about 5 to 10 mL peripheral blood by QIAamp DNA Blood midi Kit (Qiagen) and stored in −80°C for subsequent processing and analysis.

EBV antibody testing

IgA antibody titers against Epstein-Barr virus Capsid Antigen (VCA-IgA) and Early Antigen (EA-IgA) were determined from peripheral blood samples using a commercial kit based on standard immuno-enzymatic methodologies (Zhongshan Bio-tech Co. Ltd.; ref. 20), according to the manufacturer’s protocol. In brief, the protocol steps were as follows: (1) B95.8 cell smears were prepared and aliquoted in the wells of slides as antigen; (2) diluted sera were added and incubated; (3) after washing, peroxidase-conjugated antihuman IgA antibody was added and incubated; (4) wells were flooded with aminoethylcarbazole solution and H2O2; and (5) slides were then examined under the microscope. Brown staining was considered positive.

Candidate gene selection, tagging SNP selection, and genotyping

Eighty-eight genes coding for proteins with major functions in DNA repair (21) were selected as candidate genes (Table 1). SNP information was obtained from the NCBI dbSNP database and the International HapMap Project database (22). To capture the common haplotypes, tagging SNPs (htSNPs) were identified among those SNPs that were assayable (i.e., design score ≥0.60) with the GoldenGate genotyping platform (Illumina Inc.), using the aggressive multimarker selection with an r² > 0.80. A minor allele frequency (MAF) of ≥0.05 in the Han Chinese was used as a cut-off for tagging SNP selection in Haplovew software (23). A total of 676 tagging SNPs across the 88 genes were selected for genotyping.

Genotyping quality control

Manufacturer’s controls for genotyping quality were incorporated into the GoldenGate Genotyping (GGGT) assay, including allele-specific extension controls, PCR uniformity controls, gender controls, extension gap controls, first hybridisation controls, second hybridization controls, and contamination controls (http://www.lerner.ccf.org/services/gc/illumina3.php#5). In addition, we also used duplicates of 8 cases and 8 controls to test the genotyping reproducibility. The concordance rate was 100%. Moreover, a call-rate threshold of ≥95% was the criteria used to identify analyzable SNPs.

Statistical analysis

Statistical analyses were performed using R scripts (24) and PLINK software (25). The quantile–quantile (Q–Q) plot was generated using R to evaluate the overall significance of the associations of the SNPs of candidate genes. Genotypic frequencies in control subjects for each SNP were tested for
departure from Hardy-Weinberg Equilibrium (HWE) using an exact test.

Association analyses were performed using PLINK under different genetic models. In brief, the individual association of a SNP with NPC risk was evaluated by the Cochran-Armitage trend test, logistic regression, and a permutation test in PLINK. An expectation-maximization (EM) algorithm was used for haplotype imputation, reconstruction, and frequency estimations. Haplotype association tests were conducted by using the R package Haplo.stats (26). WGAViewer software (27) was used to create linkage disequilibrium (LD) plots. The criterion for statistical significance of an association was $P < 0.01$. Bonferroni correction and the Benjamini and Hochberg (28) step-up FDR approach were used for multiple comparison adjustments of the $P$ values.

Pair-wise gene–environment interactions were analyzed using generalized linear model (GLM) implemented in the R package. To assess the interactions, the likelihood ratio test was used to compare the fit of the full model (with the interaction term) and the reduced model (without the interaction term). In addition, a stratified classification by cross tabulation was also conducted. Statistic significance ($P < 0.05$) indicated a potential interaction. The pair-wised interactions between each genotyped SNP and the well-established environmental risk factors of smoking status and salted fish consumption were explored. Individuals homozygous for the major alleles and without exposures to either smoking or salted fish were used as the referent.

The classification and regression tree (CART) method implemented in the R package (rpart; ref. 29), was used to identify subgroups with higher risk for NPC classified by the genetic variants with significant marginal effects and 2 well-established risk factors. CART is a binary recursive-partitioning method that produces a decision tree to identify subgroups of subjects at higher risk. The trees were pruned and optimized into smaller ones with minimal complexity parameters. This process continued until the terminal nodes had no subsequent significant splits or the terminal nodes reached a prespecified minimum size. Five-fold cross-validation (4 of 5 of the instances for training, remaining 1 of 5 for test) was used for tree model building.

**Validation study**

To validate the significant SNPs identified in the discovery stage, we genotyped 1,568 NPC cases recruited in SYSUCC during 2 periods, from January 2002 to September 2005, and from November 2007 to March 2009, and 1,297 controls recruited in the same period from the medical examination centers of several hospitals in Guangdong. Only the subjects that were Cantonese speaking and had lived in Guangdong were included. All participants had signed informed consent and were interviewed as done before. The inclusion/exclusion criteria were the same as that described for the discovery stage. Genotyping was performed using the Sequenom DNA MassARRAY platform (Sequenom Inc.), and the procedures and analytical methods were the same as for the discovery stage.

---

**Table 1. Tagging SNP selection for the association study**

<table>
<thead>
<tr>
<th>Pathway (genes)</th>
<th>Gene symbol (genotyped SNPs/failed SNPs)$^a$</th>
<th>Subtotal of genotyped SNPs/failed SNPs</th>
<th>Successful genotyping rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER(18)</td>
<td>XRCC1(15/2), LIG3(7), APEX1(3), PARP1(7), POLE(4), OGG1(7/1), MBD4(3), NEIL2(15/1), POLD1(5), PKNP(5), SMUG1(3), UNG(5), PCNA(2), MPG(2), POLB(2/1), MUTYH(4), TDG(6), NEIL1(1)</td>
<td>96/5</td>
<td>95</td>
</tr>
<tr>
<td>HR(16)</td>
<td>NBN(10), BRCa2(20/1), RAD51(54/3), XRCC2(6), RAD54(6/1), RAD52(10), EMEl(3), RAD50(3), MRE11A(8), XRCC3(9), RAD51C(3), RAD51L3(8/2), RAD54B(6/1), BRCA1(5), RAD51(4/1), EFEMP2(1)</td>
<td>156/9</td>
<td>94</td>
</tr>
<tr>
<td>NHEJ(6)</td>
<td>XRCC4(29/4), PRKDC(11), LIG4(6/2), XRCC5(19/1), DCLRE1C(9), XRCC6(5/2)</td>
<td>79/9</td>
<td>89</td>
</tr>
<tr>
<td>NER(22)</td>
<td>RAD23B(5), RPA3(13/1), ERCC5(12), GTF2H3(4), RPA1(18), ERCC1(10/2), DDB2(5), MNAT1(15/1), GTF2H4(3), ERCC2(7), LIG1(9/1), XPC(6), RPA2(2/1), FEN1(1), ERCC4(2), XPA(6), CCNH(3/2), ERCC3(3), CDK7(4), XAB2(7), GTF2H1(6), GTF2H5(2/1)</td>
<td>143/9</td>
<td>94</td>
</tr>
<tr>
<td>MMR(5)</td>
<td>MSH3(17/1), EMS2(5/1), MLH1(4), MSH2(6), MSH6(4)</td>
<td>36/2</td>
<td>94</td>
</tr>
<tr>
<td>DR(1)</td>
<td>MGMT(43/5)</td>
<td>43/5</td>
<td>88</td>
</tr>
<tr>
<td>DDC(20)</td>
<td>RFC1(9/1), TP53(6/2), RFC4(5), HUS1(7/1), CHEK2(13), CHEK1(7/1), MDC1(5/2), MDM2(6), RAD17(3), TP53BP1(5/1), ATR(5/1), RFC3(18), CDKN1A(6/1), RAD9A(3/2), RAD1(4), ATM(5/1), RFC5(5), DPAGT1(4/1), RFC2(4), ATRIP(3/1)</td>
<td>123/15</td>
<td>88</td>
</tr>
<tr>
<td>Total (88)</td>
<td>676/54</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

$^a$The number of tagging SNPs and the number of failed SNPs are indicated in the parentheses.

Abbreviations: BER, excision repair; DDC, DNA damage checkpoints; DR, direct repair; HR, homologous recombination repair; MMR, mismatch repair; NER, nucleotide excision repair; NHEJ, nonhomologous end joining.
Table 2. NPC risk estimates of top significant SNPs in Cantonese population in Guangdong during 2005 to 2007

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Gene</th>
<th>SNP</th>
<th>Risk allele</th>
<th>Risk allele frequency (case/control)</th>
<th>OR (95% CI)</th>
<th>P_trend</th>
<th>P_emp b</th>
</tr>
</thead>
<tbody>
<tr>
<td>BER</td>
<td>OGG1</td>
<td>rs2072668</td>
<td>G</td>
<td>0.65/0.60</td>
<td>1.22 (1.04–1.43)</td>
<td>0.0068</td>
<td>0.0072</td>
</tr>
<tr>
<td>NER</td>
<td>GTF2H1</td>
<td>rs4150581</td>
<td>G</td>
<td>0.64/0.58</td>
<td>1.33 (1.13–1.57)</td>
<td>0.0006</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>MNAI1</td>
<td>rs4151400</td>
<td>C</td>
<td>0.22/0.18</td>
<td>1.23 (1.01–1.49)</td>
<td>0.0053</td>
<td>0.0043</td>
</tr>
<tr>
<td>HR</td>
<td>BRCA2</td>
<td>rs206119</td>
<td>C</td>
<td>0.18/0.14</td>
<td>1.37 (1.11–1.68)</td>
<td>0.0038</td>
<td>0.0034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs4942448</td>
<td>T</td>
<td>0.12/0.09</td>
<td>1.37 (1.07–1.76)</td>
<td>0.0072</td>
<td>0.0145</td>
</tr>
<tr>
<td>RAD51L1</td>
<td></td>
<td>rs4902562</td>
<td>G</td>
<td>0.33/0.27</td>
<td>1.27 (1.08–1.5)</td>
<td>0.0007</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs11158728</td>
<td>G</td>
<td>0.46/0.40</td>
<td>1.25 (1.07–1.45)</td>
<td>0.0009</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs927220</td>
<td>C</td>
<td>0.43/0.37</td>
<td>1.22 (1.04–1.43)</td>
<td>0.0017</td>
<td>0.0022</td>
</tr>
<tr>
<td>DDC</td>
<td>TP53BP1</td>
<td>rs689647</td>
<td>T</td>
<td>0.42/0.37</td>
<td>1.31 (1.12–1.54)</td>
<td>0.0046</td>
<td>0.0047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs544122</td>
<td>A</td>
<td>0.26/0.21</td>
<td>1.35 (1.13–1.62)</td>
<td>0.0023</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td>CHEK2</td>
<td>rs9620817</td>
<td>A</td>
<td>0.92/0.89</td>
<td>1.53 (1.16–2.00)</td>
<td>0.0037</td>
<td>0.005</td>
</tr>
</tbody>
</table>

aORs were estimated by logistic regression under the additive genetic model, adjusted by sex, age, heavy smokers, and salted fish consumption during childhood.

The empirical P values based on the statistic in Cochran-Armitage trend test. NOTE: Variants with MAF ≥ 5% and P < 0.01 are shown.

Abbreviations: BER, excision repair; DDC, DNA damage checkpoints; HR, homologous recombination repair; NER, nucleotide excision repair.

Results

Study population and genotyping success

The average age of the cases was 46.6 years old, and the sex ratio was 2.8 males to 1 female, suggesting the case representation paralleled the demographics of NPC in the general population (Supplementary Table S1). Of the 676 tagging SNPs among 88 genes (Table 1), 54 SNPs (8%) were eliminated either due to poor genotyping (call rate < 0.95) or significant deviation from HWE (P < 0.01). The remaining 622 SNPs (92%) were included in analyses of NPC risk. A quantile–quantile (Q–Q) plot revealed a good match between the distributions of the observed P values and those expected by chance, except for deviation within the tail of the distribution.

Variants of DNA repair genes and NPC risk

Eleven SNPs among 7 DNA repair genes (RAD51L1, BRCA2, TP53BP1, OGG1, MNAI1, CHEK2, and GTF2H1) showed statistically significant associations with NPC (P < 0.01; Table 2; Supplementary Fig. S1). Three genes—RAD51L1, BRCA2, and TP53BP1—had multiple significant loci.

Three neighboring SNPs in RAD51L1 exhibited strong association with NPC risk (rs4902562: P_trend = 0.0007, rs11158728: P_trend = 0.0009, rs927220: P_trend = 0.0017). $r^2$ for the rs4902562-rs11158728 pair was 0.41 ($D' = 0.85$), for rs4902562-rs927220 pair was 0.35 ($D' = 0.74$), for rs11158728-rs927220 was 0.81 ($D' = 0.94$). An additional 4 loci in this gene showed weak associations with NPC risk (rs4899234: P_trend = 0.0217, rs6573824: P_trend = 0.0256, rs911263: P_trend = 0.0242, rs7148412: P_trend = 0.0236, respectively). After partitioning RAD51L1 into 11 LD blocks using Gabriel’s Confidence Intervals method, haplotype association with NPC was analyzed for each LD block. Two haplotypes showed significant association with an increased risk for NPC. Haplotype GTG in block #4, partitioned by rs4902562, rs1957572, and rs6573824, had a P value of 0.0005 (OR = 1.26; 95% CI = 1.06–1.50). In addition, the haplotype TGC in block #5, partitioned by rs12432392, rs11158728, and rs927220, also had a significant association (OR = 1.29; 95% CI = 1.11–1.49; $P = 0.0017$; Supplementary Table S2). A fixed length “sliding” technique using 3 consecutive SNPs to scan the effects of haplotypes across RAD51L1 indicated at least 2 peaks with strong signals of association (Fig. 1).

BRCA2 was also shown to be significantly associated with NPC at 2 SNPs (rs206119: $P_{trend} = 0.0038$; rs4942448: $P_{trend} = 0.0022$; Table 2). Furthermore, significant associations for TP53BP1 were detected for 2 SNPs (rs689647: $P_{trend} = 0.0046$; rs544122: $P_{trend} = 0.0023$; Table 2). Haplotype-based association analyses further supported the association of BRCA2 and TP53BP1 with NPC risk (Supplemental Table S2).

In addition to RAD51L1, BRCA2, and TP53BP1, which had multiple SNPs associated with NPC, the OGG1, GTF2H1, MNAI1, and CHEK2 genes each had a single SNP variant associated with NPC. The risk allele “G” of rs2072668 in OGG1 had $P_{trend} = 0.0068$; the risk allele “G” of rs4150581 in GTF2H1 had $P_{trend} = 0.0006$; the risk allele “C” of rs4151400 in MNAI1 had $P_{trend} = 0.0053$; and the risk allele “A” of rs9620817 in CHEK2 had $P_{trend} = 0.0037$ (Table 2). Permutation-based $P$ values also supported associations with NPC (Table 2). The 11 SNPs all had permutation test $P < 0.01$, except for rs4942448 which had a P value of 0.0145.

Exploring gene–environment interaction in susceptibility to NPC

In our study, 96% of NPC cases were EBV VCA-IgA positive, compared with only 19% of controls. After adjusting by age, sex,
and education, heavy smokers had a moderate risk with OR of 1.81 (95% CI 1.39–2.35), and the individuals who consumed salted fish during childhood had an OR of 3.28 (95% CI 2.64–4.09).

GLM pair-wise interaction analyses suggested that "consumption of salted fish during childhood" had interactions with variants of DNA repair genes TP53BP1, RAD50, RAD54B, RAD54L, LIG1, LIG3, LIG4, and POLE, with increased ORs ranging from 3.16 to 4.43 compared with those individuals harboring the homozygous major alleles and without exposure (the referent). Similarly, smoking status appeared to

Figure 1. Gene structure of RAD51L1, linkage disequilibrium patterns, and results of haplotype and SNP association with NPC risk. A, the ideogram of chromosome 14, indicating the position of RAD51L1 and its gene context (1 of 3 mRNA variants is shown, vertical lines indicate the exons). B, haplotype association scanning was conducted by Haplo.stat. In brief, to evaluate the association of sub-haplotypes (subsets of alleles from the full haplotype) with NPC risk, we evaluated a "window" of alleles by sliding a fixed-width window (3 SNPs) across the entire haplotype in a gene. The y-axis depicts P values on a minus logarithmic scale; the x-axis represents the SNPs ordered along the chromosome according to their positions. The blue arrows indicate the peaks of negative logarithm of P values (−log P) of "fixed-width window" haplotype association. Note that this method used 3 consecutive SNPs to evaluate the subhaplotype score and test the significance; the 3 SNPs (the subhaplotype) will have a single P value. It produces multiple horizontal lines when plotting the −log P versus the SNPs alongside the x-axis. C, P values for single SNP association tests (the 1-df Cochran-Armitage trend test) in RAD51L1 gene. Red vertical lines indicate significant SNPs with Ptrend < 0.01. Their positions on the linkage disequilibrium diagram are indicated by black squares. D, the estimates of linkage disequilibrium parameter (R-Square). Colors are coded according to the scale and shown at the bottom left.
interact with genetic variants of *MGMT*, *RFC3*, *RAD51L1*, and *RAD52*, with ORs ranging from 1.52 to 1.82 (Supplementary Table S5). Due to the fact that almost all cases were EBV positive (i.e., 96% positive by VCA-IgA titer), we did not have adequate power to identify any genotypic interaction with EBV exposure.

The CART method identified a few subgroups with higher risk of NPC among the risk factors and the genetic variants of the *TP53BP1*, *RAD51L1*, *BRCA2*, and *MNAT1* genes (Fig. 2; Supplementary Table S6). There was an initial split on salted fish consumption, suggesting that salted fish consumption was one of the most important risk factors for NPC. More interestingly, distinct patterns for consumption and nonconsumption were observed in the CART tree. That is, in the right branch of the tree, the nodes with consumption of salted fish showed higher ORs than the nonconsumption nodes in the left branch. As compared with the referent (Node #5), those individuals who consumed salted fish and carried allelic variants showed dramatically increased risk of NPC (e.g., those carriers of variant rs544122 in *TP53BP1* were at the highest risk (OR = 4.13 (95% CI = 2.41–7.17), 5.59 (95% CI = 2.59–12.60), and 6.42 (95% CI = 3.83–10.96), respectively. Moreover, a subgroup with higher risk was also identified in the salted fish nonconsumption populations. In node #16, the heavy smokers showed higher risk to NPC when carrying the variant allele of rs11158728, OR = 3.39 (95% CI = 2.04–5.70). Nevertheless, in node #12, although nonheavy smokers, the populations carrying multiple variant alleles still showed a higher risk, OR = 2.72 (95% CI = 1.53–4.84; Fig. 2; Supplementary Table S6).

However, when we tested the potential interactions between rs544122 and salted fish consumption during childhood by comparing the full model \( z = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 \) to the reduced model \( z = b_0 + b_1 x_1 + b_2 x_2 \), we did not observe any interaction between these 2 variables \( P = 0.3076 \).

**Validation of the significant SNPs in a separate population**

Table 3 shows the results of the analyses of the 11 most significantly associated SNPs in the validation set. rs927220 in *RAD51L1* was significantly associated with NPC risk \( P_{\text{trend}} = 0.0085 \) in the validation data set alone and \( P_{\text{trend}} = 5.55 \times 10^{-5} \).
in the combined data set). In addition, rs11158728 in RAD51L1 provided a weak association ($P_{\text{trend}} = 0.0335$ for the validation data set and $P_{\text{trend}} = 0.0002$ for the combined data set; see the Table 3). Moreover, rs544122 in TP53BP1 and rs4942448 in BRCA2 had $P_{\text{trend}} < 0.05$ either in the validation set or in the combined data set.

We then used the Bonferroni correction and the Benjamini and Hochberg step-up FDR approach for multiple comparison adjustments. In particular, SNP rs927220 suggested evidence of association after being corrected for multiple testing (Bonferroni correction). SNP rs927220 in RAD51L1 was significant after multiple comparison correction ($P = 0.0381$ after Bonferroni correction). In addition, another SNP in RAD51L1 (rs11158728) also had a significant combined $P$ value ($2.0 \times 10^{-5}$), but fell below significance after multiple comparison correction. These 2 SNPs are in strong LD with each other ($\chi^2 = 10.0, r^2 = 0.7$).

Further, we explored the LD patterns between SNPs with possible function in RAD51L1 and the significantly associated SNPs (rs11158728, rs4902562, and rs927220), but no putative functional SNPs have yet been identified as being in strong LD with these SNPs. A total of 53 putative functional SNPs were identified using the latest dbSNP (Build 131) based on their gene locations, including 17 SNPs within the coding region, 30 SNPs located within 2 kb upstream of the gene, 1 SNP located within 0.5 kb downstream of the gene and 5 SNPs in the 3' UTR. They are considered "potential functional SNPs". Using the Chinese Han population (CHB) genotype data within HapMap, we found that only 2 out of 53 SNPs had qualified genotypes, and LD between the 3 NPC-associated SNPs and 2 potential functional SNPs was low (Supplementary Table S3). In contrast, the SNPs in strong LD ($r^2 \geq 0.5$) with rs11158728, rs4902562, and rs927220 in RAD51L1 are all located within introns (data not shown). Further investigations need to be conducted to make conclusive comments about the relationship of these SNPs and the putative functional SNPs.

RAD51L1 is a member of the RAD51 protein family, which regulates the central activity of homologous recombination (HR) DNA repair (30). RAD51 and its homologues have been reported to be very important to cancer risk, particularly for breast and ovarian cancers (31). Recently, a 3-stage GWAS on breast cancer in 9,770 cases and 10,799 controls mapped the susceptibility locus to RAD51L1 (32). RAD51L1 has been

---

**Table 3. Validation in a separate Cantonese population recruited in Guangdong during 2002 to 2005 and 2007 to 2009 for the 11 top significant SNPs in the discovery stage**

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Risk allele</th>
<th>$P_{\text{hwe}}$/Call rate (%)</th>
<th>Risk allele frequency (case/control)</th>
<th>OR (95% CI)</th>
<th>$P_{\text{trend}}$</th>
<th>Risk allele frequency (case/control)</th>
<th>OR (95% CI)</th>
<th>$P_{\text{trend}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGG1</td>
<td>rs2072668</td>
<td>G</td>
<td>0.0599/99.5 0.60/0.60</td>
<td>0.99 (0.88–1.10) 0.7531</td>
<td>0.02/0.01</td>
<td>1.06 (0.97–1.16) 0.2051</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTF2H1</td>
<td>rs4150581</td>
<td>G</td>
<td>0.4899/99.9 0.60/0.60</td>
<td>0.98 (0.88–1.10) 0.8204</td>
<td>0.01/0.01</td>
<td>1.09 (0.90–1.25) 0.0701</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNAT1</td>
<td>rs4151400</td>
<td>C</td>
<td>0.5415/99.5 0.20/0.19</td>
<td>1.06 (0.92–1.21) 0.3527</td>
<td>0.21/0.19</td>
<td>1.13 (1.02–1.26) 0.0173</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>rs206119</td>
<td>C</td>
<td>0.4885/99.6 0.17/0.17</td>
<td>0.99 (0.86–1.14) 0.7296</td>
<td>0.17/0.16</td>
<td>1.08 (0.86–1.21) 0.1375</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>rs4942448</td>
<td>T</td>
<td>0.0858/99.9 0.10/0.11</td>
<td>0.97 (0.82–1.16) 0.0369</td>
<td>0.10/0.10</td>
<td>1.08 (0.84–1.24) 0.2400</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD51L1</td>
<td>rs902562</td>
<td>G</td>
<td>0.0099/99.9 0.31/0.32</td>
<td>0.97 (0.86–1.08) 0.6289</td>
<td>0.32/0.30</td>
<td>1.09 (0.99–1.19) 0.0923</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD51L1</td>
<td>rs11158728</td>
<td>G</td>
<td>0.0005/99.5 0.45/0.42</td>
<td>1.12 (1.01–1.24) 0.0335</td>
<td>0.45/0.41</td>
<td>1.17 (1.08–1.27) 0.0002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD51L1</td>
<td>rs927220</td>
<td>C</td>
<td>0.0755/98.5 0.43/0.40</td>
<td>1.14 (1.03–1.27) 0.0085</td>
<td>0.43/0.39</td>
<td>1.20 (1.10–1.30) 5.55 × 10^{-5}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53BP1</td>
<td>rs689647</td>
<td>T</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53BP1</td>
<td>rs544122</td>
<td>A</td>
<td>0.3444/99.3 0.25/0.24</td>
<td>1.05 (0.93–1.19) 0.3234</td>
<td>0.26/0.23</td>
<td>1.14 (1.03–1.26) 0.0080</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEK2</td>
<td>rs9620817</td>
<td>A</td>
<td>Failed</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $n$, the number of cases vs. controls.
- NA, the SNP could not be subjected to the Sequenom genotyping platform.
- Odds ratio was adjusted by sex and age.

A haplotyping-tagging approach was used to systematically investigate the associations between hundreds of common variants of DNA-repair genes and NPC risk. The advantage of this comprehensive approach is that it covered all of the major human DNA repair pathways and virtually all known human DNA repair genes. This is, thus far, the most comprehensive multigenic study evaluating associations between a large number of DNA repair gene variants and NPC.

In the discovery stage, the most notable finding was that 7 of the 11 most significant variants were located within just 3 genes, RAD51L1, BRCA2, and TP53BP1. These associations were maintained at both the individual SNP and at the haplotype levels.

In the validation stage, within a separate Cantonese population, 2 of the SNPs, both located in RAD51L1, retained their association with NPC. Specifically, rs927220 in RAD51L1 was the most strongly associated, with a combined $P = 5.55 \times 10^{-5}$, and remained significant after multiple comparison correction ($P = 0.0381$ after Bonferroni correction). In addition, another SNP in RAD51L1 (rs11158728) also had a significant combined $P$ value ($2.0 \times 10^{-5}$), but fell below significance after multiple comparison correction. These 2 SNPs are in strong LD with each other ($\chi^2 = 10.0, r^2 = 0.7$).
reported to form a stable heterodimer with RAD51 family member RAD51L2, which then further interacts with the other family members, such as RAD51, as well as XRCC2 and XRCC3 (33). RAD51L1(−/−) cells are hypersensitive to DNA cross-linking agents and have much reduced formation of RAD51 nuclear foci (34), suggesting that DNA repair is compromised when the gene is dysfunctional.

In addition to RAD51L1, we found that BRCA2 was also associated with NPC in the discovery stage. BRCA2 is another major gene involved in the HR repair pathway and the BRCA2 protein physically interacts with RAD51 proteins (35), and thus has functional relevance to RAD51L1. Nevertheless, the BRCA2 SNPs failed to validate.

Previous studies have reported that OGG1 was associated with NPC (36, 37) and other cancers (38, 39). In our study, rs2072668 in OGG1 showed statistically significant evidence of association in the discovery stage (P = 0.0068). This SNP is in strong LD with nonsynonymous SNP rs1052133 (r² = 1.0; Supplementary Table S3) that may be functional. Yet, rs2072668 was not significantly associated with NPC in the validation stage.

Because NPC is a malignancy that fits a multifactorial model, it is important to account for environmental risk factors when assessing genetic associations. Pair-wise interaction results suggested possible interactions between environmental risks and genetic components. As indicated in Table S5, many of the variants involved in the putative interactions were from the RAD gene family (i.e., RAD50, RAD54B, RAD54L, and RAD51L1). Although the pattern of genes with potential interactions shows a degree of commonality, some displayed negative interactions, which are hard to interpret mechanistically. This raises some question as to whether the statistical interactions found here have biological relevance.

CART models yielded large ORs (ranging from 4.1 to 7.6) for individuals consuming salted fish and carrying more than 1 risk genotype in RAD51L1, TP53BP1, and BRCA2. Statistically significant departures from multiplicative models were not observed; however, much larger sample sizes will be needed to assess the joint effects of DNA repair variation and nongenetic risk factors of NPC.

Previous studies have identified other loci associated with NPC. Among them, HLA has consistently been reported as a susceptibility region (40–42). [Other susceptibility loci include 4p15.1-1q2 (43) and 3p21 (44).] Recently, 2 independent genome-wide association studies (GWAS) supported the HLA region as an NPC risk locus in the Cantonese (45) and Taiwanese (46). The third relatively small-scale GWAS found that the integrin-alpha 9 gene (ITGA9) was associated with NPC risk (47).

Our candidate gene association study provided a focused view of specific DNA repair genomic regions. Only 1 SNP in RAD51L1 (rs927220) survived the Bonferroni correction. This same SNP had also been genotyped in a previous GWAS study (45) and had a moderate P value of 0.02 in Cantonese population. The P value below the threshold of our GWAS phase I study for further validation study might due to its lack of power. More importantly, the directions of the association effect revealed by the previous GWAS and current candidate gene study are consistent [OR in GWAS is 1.17 (95% CI = 1.02–1.35); OR in candidate gene study is 1.22 (95% CI = 1.04–1.43)]. We believed that this candidate gene study could be a complement to the findings from previous GWAS.

In summary, rs927220 in RAD51L1 was discovered to be associated with NPC, and subsequently validated at a significant level in another study group. This SNP is located in an intron, and therefore is unlikely to be functional itself, but it may be in LD with a yet to be identified functional locus. Other SNP associations identified in the discovery stage could not be validated and further studies are needed. This study supports the notion that DNA repair genes, in particular the RAD51L1 gene, might play a role in NPC etiology and development.

Disclosure of Potential Conflicts of Interest

The authors declared no conflict of interest.

Acknowledgments

We gratefully acknowledge the participation of all individuals, without whom this work would not have been possible. We also thank Professor Shao-Qi Rao for his help in manuscript revision.

Grant Support

This work was supported by the National Natural Science Foundation of China (30671798, 30471487, u0732005), the National Science and Technology Support Program of China (2006BAI02A11), the National Major Basic Research Program of China (863: 2006AA02A404) and the National Institute of Health (ROCA113240–01), USA.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 5, 2010; revised January 5, 2011; accepted February 17, 2011; published OnlineFirst March 2, 2011.

References


Comprehensive Pathway-Based Association Study of DNA Repair Gene Variants and the Risk of Nasopharyngeal Carcinoma

Hai-De Qin, Yin Yao Shugart, Jin-Xin Bei, et al.

Cancer Res Published OnlineFirst March 2, 2011.