Journal of Human Genetics (2013) 58, 679–685

www.nature.com/jhg

ORIGINAL ARTICLE

Association of *XRCC3* and *XRCC4* gene polymorphisms, family history of cancer and tobacco smoking with non-small-cell lung cancer in a Chinese population: a case–control study

Fei He[1], Shen-Chih Chang[2], Gina Maria Wallar[2], Zuo-Feng Zhang[2] and Lin Cai[1]

Single-nucleotide polymorphisms (SNPs) of DNA repair genes have been reported to modify cancer risk. This study aimed to determine SNPs of the DNA repair genes X-ray repair cross-complementing group 3 (*XRCC3*) and X-ray cross-complementing group 4 (*XRCC4*) and their association with non-small-cell lung cancer (NSCLC) susceptibility in a Chinese population. A total of 507 NSCLC patients and 662 healthy controls were recruited for genotyping. Epidemiological and clinical data were also collected for association studies. The data showed that the rs1799794 G allele in the *XRCC3* gene and minor allele carriers of *XRCC4*, including rs1056503 and rs9293337, were inversely associated with NSCLC risk (GG vs homozygote AA), whereas the rs861537 AG or AA genotype and *XRCC4* rs6869366 had a significantly increased NSCLC risk. Furthermore, tobacco smoking over 26 pack-years, a family history of lung cancer, exposure to environmental tobacco smoke (ETS) and negative mental status were risk factors for developing NSCLC. This study suggests that SNPs of *XRCC3* and *XRCC4* and other environmental factors are risk factors for developing NSCLC in this Chinese Han population.
Journal of Human Genetics (2013) **58,** 679–685; doi:10.1038/jhg.2013.78; published online 8 August 2013

Keywords: case–control study; non-small-cell lung cancer; single-nucleotide polymorphism; *XRCC3* and *XRCC4*

INTRODUCTION

Lung cancer was the most commonly diagnosed cancer and the leading cause of cancer-related deaths in men and was the fourth most commonly diagnosed cancer and the second leading cause of cancer-related deaths in women in 2008, accounting for approximately 1.6 million worldwide new cases and 1.4 million deaths in 2008.[1] Histologically, lung cancer can be divided into small-cell lung cancer and non-small-cell lung cancer (NSCLC). The latter includes squamous cell carcinoma, adenocarcinoma and large-cell carcinoma. Survival of lung cancer patients is still very low; thus, prevention and early detection could reduce the burdens of lung cancer.

Tobacco smoke is an important risk factor for developing lung cancer, accounting for 80% of male lung cancers and 50% of female cases. Studies of familial aggregation in lung cancer suggest that inherited genetic susceptibility may contribute to tumorigenesis.[2] Tobacco smoke can induce DNA damage,[3] such as DNA double-strand breaks (DSBs), which are the most detrimental form of DNA damage and can result in gene mutation, cell death or neoplastic transformation. In addition, improperly repaired DSBs can induce a high predisposition to gene translocations and tumorigenesis. To date, there are two types of DNA DSB repair processes, including homologous recombination and non-homologous end-joining DNA repairs. Homologous recombination promotes accurate repair of DSBs by copying intact information from an undamaged homologous DNA template. Non-homologous end-joining is a homology-independent mechanism that rejoins broken ends irrespective of DNA sequences.[4,5] The X-ray repair cross-complementing group 3 (*XRCC3*) gene belongs to the homologous recombination pathway and its protein helps to maintain chromosome stability and repair damaged DNA when DSBs occur. In addition, it has an essential role in maintaining chromosome stability during cell division.[6,7] XRCC3 is also a member of the RecA/Rad51 family of proteins that includes seven recA-like genes, including *RAD51*, *RAD51L1/B*, *RAD51L2/C*, *RAD51L3/D/TRAD*, *XRCC2*, *XRCC3* and *DMC1*.[8–10] Moreover, the X-ray cross-complementing group 4 (*XRCC4*) gene, an important component of non-homologous end-joining, works in conjunction with Ku70/Ku80 and ligase 4 to have a major role in the precision end-joining of blunt DSBs.[11] A study using animal models showed that mouse embryonic cells with disrupted *XRCC4* had reduced proliferation, radiation hypersensitivity, chromosomal instability and severely impaired variable (V), diversity (D) and joining (J) (V(D)J)

[1]Department of Epidemiology, School of Public Health, Fujian Medical University, Fuzhou, China and [2]Department of Epidemiology, School of Public Health, University of California, Los Angeles (UCLA), Los Angeles, CA, USA
Correspondence: Dr L Cai, Department of Epidemiology, School of Public Health, Fujian Medical University, Fuzhou Jiaotong Road, Fuzhou 350004, China.
E-mail: Cailin_cn@hotmail.com
Received 4 March 2013; revised 4 June 2013; accepted 3 July 2013; published online 8 August 2013

recombination.[12] These data indicate that alteration of DNA repair genes is associated with human carcinogenesis.

Thus, single-nucleotide polymorphisms (SNPs) in the *XRCC3* and *XRCC4* genes may contribute to the unrepaired DNA damage in the human genome, resulting in susceptibility to tumorigenesis.[13–15] Some studies have identified polymorphisms in *XRCC3* and *XRCC4* to be associated with lung cancer risk, including rs861539,[16] rs6869366[17] and rs1799796.[18] However, the results of some studies[19,20] on lung cancer or other cancer type were inconsistent. A few other studies focused on NSCLC susceptibility in association with *XRCC3* and *XRCC4*.[21,22] In this study, we investigated the associations between NSCLC and SNPs of *XRCC3* (rs861539, rs1799796, rs861537 and rs1799794) and *XRCC4* (rs9293337, rs6869366, rs3734091 and rs1056503) in a Chinese Han population. We also examined the joint effects of haplotypes and diplotypes of *XRCC3* and *XRCC4* and their association, together with individual social behavior factors, with NSCLC risk.

MATERIALS AND METHODS

Study subjects

In this study, we recruited 507 patients with newly diagnosed primary NSCLC and 662 controls subjects from our three area hospitals (The First Clinical Medical College of Fujian Medical University, The Affiliated Union Hospital of Fujian Medical University and Fuzhou General Hospital) between July 2006 and September 2009. Second primary and recurrent NSCLC cases were excluded from this study. The recruiting rate for NSCLC patients was 96.93% and the rate for control subjects was 92.0%. All cases and controls resided in Fuzhou City or in the surrounding regions of Fujian Province. The controls were randomly selected from the community and matched to the cases by age (± 3 years) and gender. Individuals who were direct relatives to the cases or had a previous history of cancer were excluded. This case–control study was approved by the Institutional Review Board (IRB) of Fujian Medical University (Fuzhou, China) and University of California at Los Angeles (UCLA, USA). All participants agreed to this study and signed a consent form.

Data and sample collections

All epidemiological data were obtained by in-person interviews using a standardized questionnaire, which collected information on demographic characteristics, socioeconomic status, diet, family history of lung cancer and living environment. Furthermore, history of tobacco use and exposure was collected to determine the age when the patient started to smoke, as well as the duration, amount and levels of exposure to environmental tobacco smoke (ETS).

Smokers were defined as individuals who had smoked at least 100 cigarettes during their lifetime. Cumulative smoking was quantified by pack-years ((cigarettes per day/20) $\times$ (years smoked)). Light and heavy smokers were categorized using the 50th percentile of pack-years in the controls. ETS was defined as exposure to other ETS at home and/or at work for more than 15 min per day. Mental status was defined in two states, pessimistic as negative and optimistic as positive.

A 5-ml non-fasting blood sample was collected from each case with a vacuum tube. A saliva sample was also collected from each control using the Oragene DNA Self-Collection Kit (DNA Genotek, Ottawa, ON, Canada).

Selection of SNPs

In this study, based on the HapMap-CHB genotype data (HapMap Data Rel 27 Phase II + III, Feb09, on NCBI B36 assembly, dbSNP b126), we set an r^2 threshold of 0.5 and a minor allele frequency >0.10 as suggested by the study by Carlson *et al.*[23] *XRCC3* tSNPs rs861537 and rs1799794 were identified using a haplotype-based tagging method by de Bakker *et al.*[24] in Haploview program. Considering the inconsequent results in the previous studies, rs861539 and rs1799796 in *XRCC3* were selected as candidate SNPs. For *XRCC4*, tSNPs rs3734091 and rs1056503 were in the coding SNP region and rs6869366 was in the promoter region. According to the sequence-based approach,[25] these three SNPs may be the underlying functional SNPs. And, we chose rs9293337 as the candidate SNP in lung cancer first time.

Genotyping

Genomic DNA from the blood and saliva samples was extracted using a protease K digestion and phenol–chloroform extraction and purification according to a standard operation procedure. The genomic DNA was then stored at $-20\,^{\circ}C$. After that, DNA samples were subjected to SNP genotyping using the Sequenom platform according to the manufacturer's iPLEX Application Guide (Sequenom, San Diego, CA, USA). Polymerase chain reaction (PCR) and extension primers were designed using MassARRAY Assay Design 3.1 software (Sequenom). Briefly, PCR amplification was conducted in a total volume of 5 μl with 10 ng of genomic DNA, 3.5 mM of $MgCl_2$, 0.5 U of HotStarTaq polymerase (Qiagen, Valencia, CA, USA), 500 μM of dNTP (Invitrogen, Carlsbad, CA, USA) and 60 nM of each primer set. The PCR conditions were as follows: 94 °C for 15 min followed by 45 cycles of 94 °C for 20 s, 56 °C for 30 s and 72 °C for 1 min, and a final extension at 72 °C for 3 min. The PCR products were subjected to shrimp alkaline phosphatase treatment in a total volume of 7 μl with 0.3 U of shrimp alkaline phosphatase enzyme and then incubated at 37 °C for 40 min. Next, the products were further subjected to iPLEX reactions in a total volume of 9 μl with 1 $\times$ iPLEX termination mix, 1 $\times$ iPLEX enzyme and 5.625 μM of each extension primer. The products were incubated at 94 °C for 30 s followed by a total of 200 nested PCR cycles consisting of 40 main cycles at 94 °C for 5 s, five subcycles at 52 °C for 5 s and 80 °C for 5 s, and a final extension at 72 °C for 3 min. The products were then cleaned with 6 mg of resin and applied to a SpectroCHIP. At the end of the experiments, the samples were scanned through a MALDI-TOF MS system and genotypes were analyzed by the MassArrayTyper 3.4 (Sequenom). Approximately 10% of the samples were randomly repeated for quality control purposes. Genotyping call rates were $>94.0\%$ and the concordance rate reached 99.5%.

Statistical analysis

Statistical software PASW v.19.0 (IBM Corporation, Somers, NY, USA) was used for our data analyses. Two-sided χ^2 tests were performed to compare differences in distributions of selected demographic factors between cases and controls. Unconditional logistic regression models were used to estimate odds ratios (ORs) and their 95% confidence intervals (CIs). Potential confounders were selected based on prior knowledge of lung cancer, which include age, gender, education, family history of lung cancer, smoking status, ETS, and mental and marital status. Hardy–Weinberg equilibrium was conducted using a goodness-of-fit χ^2 test with linkage disequilibrium analyzer (LDA) software v.1.0 (Chinese National Human Genome Center, Beijing, China) for each SNP among the controls. Haplotypes and diplotypes of *XRCC3* and *XRCC4* and their associations with NSCLC were determined using PHASE 2.1 software (Department of Statistics, University of Chicago, Chicago, IL, USA). Gene–environment interactions were determined using the association rule mining method with SPSS Clementine v.12.0 (IBM Corporation). The variables that were significant in both univariate and association analyses were subjected to logistic multifactor analysis.

We used association rule mining to identify the strong associations that satisfied predefined minimum support and confidence at the same time from a given database. Thus, we first searched item sets called frequent or large item sets whose occurrences exceeded a predefined threshold in the database. We then generated the association rules from the frequent item sets with the constraints of minimal confidence or deleting the last items in the antecedent. We inserted these sets to the consequent and then determined the interests of the rules by checking the confidence levels. These processes iterated until the antecedent became empty. Next, the three most important evaluating indicators were defined as the support, confidence and lift. The lift was defined as the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence was the product of the support values of the rule body and the rule head divided by the support of the rule body. The confidence value was the ratio of the support of the joined rule body and rule head divided by the support of the rule body. Clementine was the data-mining workbench that offered a number of algorithms for

clustering, classification, association and prediction, as well as algorithms for automated multiple modeling, time-series forecasting and interactive rule building. These algorithms exist in a Clementine 'base' module with optional additional modules. In this study, an apriority algorithm was used to analyze the data stream to identify the association between the variants and disease outcome.

RESULTS

Characteristics of study subjects

In this study, we recruited a total of 507 cases and 662 controls. Demographic characteristics and risk factors between cases and controls are described in Table 1. In particular, tobacco smoke, especially heavy smoking and starting to smoke earlier in life, exposure to ETS, family cancer history and a negative mental status were associated with NSCLC. In contrast, a high educational background was a protective factor for NSCLC. People with a negative mental status had an increased risk for developing NSCLC with or without a history of tobacco smoking ($P = 0.001$ and 0.010, respectively). A family history of lung cancer was also associated with NSCLC (odds ratio (OR) = 1.47; 95% confidence interval (CI): 1.09–1.98), especially among tobacco smokers (OR = 1.94; 95% CI: 1.21–3.12).

Individual SNP and haplotype effects on NSCLC risk

We then detected the individual SNP and haplotype effects of *XRCC3* and *XRCC4* on NSCLC risk. Data in Table 2 show the genotype

Table 1 Distribution of selected variables among cases and controls in Han population

Variables (n = *1169*)	*Case (%)* (n = *507*)	*Control (%)* (n = *662*)	P-*value*	*Odds ratios (95% confidence intervals)*
Age (years), mean ± s.d.	58.91 ± 11.55	58.84 ± 11.95	0.621	
⩽50	117 (23.1)	154 (23.3)		1
51–69	296 (58.4)	371 (56.0)		1.05 (0.79–1.40)
⩾70	94 (18.5)	137 (20.7)		0.90 (0.63–1.29)
Gender			0.827	
Male	369 (72.8)	478 (72.2)		1
Female	138 (27.2)	184 (27.8)		1.03 (0.79–1.33)
Education			**0.001**	
Illiteracy	71 (14.0)	64 (9.7)		1
Middle school and below	303 (59.8)	359 (54.2)		0.76 (0.53–1.10)
High school and above	133 (26.2)	239 (36.1)		**0.50 (0.34–0.75)**
Family history of lung cancer			**0.011**	
No	397 (78.3)	557 (84.1)		1
Yes	110 (21.7)	105 (15.9)		**1.47 (1.09–1.98)**
ETS			**<0.001**	
No	123 (24.3)	436 (65.9)		1
Yes	384 (75.7)	226 (34.1)		**6.02 (4.65–7.80)**
Marital status			0.272	
Married	477 (94.1)	612 (92.4)		1
Single	30 (5.9)	50 (7.6)		0.77 (0.48–1.23)
Mental status			**0.001**	
Positive	410 (80.9)	582 (87.9)		1
Negative	97 (19.1)	80 (12.1)		**1.72 (1.25–2.38)**
Age start smoke (n = *565*)			**0.001**	
⩽20	229 (71.6)	141 (57.6)		1
⩾21	91 (28.4)	104 (42.4)		**0.54 (0.38–0.77)**
Smoking pack-years (mean ± s.d.)	10.82 ± 19.01	31.40 ± 34.88	**<0.001**	
Never	187 (36.9)	416 (63.2)		1
⩽25	70 (13.8)	125 (18.9)		1.25 (0.89–1.75)
⩾26	250 (49.3)	118 (17.9)		**4.71 (3.57–6.23)**
Histology				
Adenocarcinoma	245 (48.3)			
Squamous cell carcinoma	181 (35.7)			
Others	81 (16.0)			

Abbreviation: ETS, environmental tobacco smoke.
Bold numerals: A P-value of <0.05 was considered statistically significant.

Table 2 Distribution of *XRCC3* and *XRCC4* single-nucleotide polymorphisms and their associations with non-small-cell lung cancer in Han population

Locus	*Case (%) (n = 507)*	*Control (%) (n = 662)*	*Unadjusted odds ratios (95% confidence intervals)*	*Adjusted odds ratios* (95% confidence intervals)*	P_{trend}*-value*
XRCC3					
rs861539 ($P_{HWE} = 0.99$)	507	660			0.692
CC	450 (88.8%)	589 (89.2%)	1	1	
CT	54 (10.7%)	69 (10.5%)	1.02 (0.70–1.49)	1.12 (0.83–1.51)	
TT	3 (0.5%)	2 (0.3%)	1.96 (0.33–11.80)	1.10 (0.73–1.65)	
CT + TT	57 (11.2%)	71 (10.8%)	1.05 (0.73–1.52)	1.10 (0.72–1.68)	
rs1799796 ($P_{HWE} = 0.07$)	455	652			0.175
AA	174 (34.3%)	276 (42.3%)	1	1	
AG	206 (40.6%)	280 (42.9%)	1.17 (0.90–1.52)	1.11 (0.82–1.49)	
GG	75 (34.3%)	96 (14.8%)	1.24 (0.87–1.77)	1.09 (0.73–1.64)	
AG + GG	281 (61.8%)	376 (57.7%)	1.19 (0.93–1.51)	1.10 (0.83–1.46)	
rs861537 ($P_{HWE} = 0.16$)	507	660			**0.001**
GG	147 (29.0%)	254 (38.5%)	1	1	
AG	255 (50.3%)	297 (45.0%)	**1.48 (1.14–1.93)**	**1.44 (1.06–1.94)**	
AA	105 (20.7%)	109 (16.5%)	**1.66 (1.19–2.33)**	**1.53 (1.04–2.25)**	
AG + AA	360 (71.0%)	406 (61.5%)	**1.53 (1.20–1.96)**	**1.46 (1.10–1.94)**	
rs1799794 ($P_{HWE} = 0.18$)	507	661			**0.002**
AA	180 (35.5%)	184 (27.8%)	1	1	
AG	230 (45.4%)	313 (47.4%)	**0.75 (0.58–0.98)**	0.80 (0.59–1.09)	
GG	97 (19.1%)	164 (24.8%)	**0.61 (0.44–0.84)**	**0.60 (0.41–0.87)**	
AG + GG	327 (64.5%)	477 (72.2%)	**0.70 (0.55–0.90)**	**0.73 (0.55–0.98)**	
XRCC4					
rs6869366 ($P_{HWE} = 0.94$)	501	651			**0.009**
TT	437 (87.2%)	599 (92.1%)	1	1	
TG	63 (12.6%)	51 (7.8%)	**1.69 (1.15–2.50)**	**1.85 (1.17–2.92)**	
GG	1 (0.2%)	1 (0.1%)	1.37 (0.09–21.97)	2.12 (0.12–36.72)	
TG + GG	64 (12.8%)	52 (7.9%)	**1.69 (1.15–2.48)**	**1.86 (1.18–2.91)**	
rs3734091 ($P_{HWE} = 0.14$)	505	654			0.397
CC	443 (87.7%)	583 (89.1%)	1	1	
CA	61 (12.1%)	71 (10.9%)	1.13 (0.79–1.63)	1.08 (0.71–1.65)	
AA	1 (0.2%)	0 (0.0%)	–	–	
CA + AA	62 (12.3%)	71 (10.9%)	1.15 (0.80–1.65)	1.12 (0.73–1.69)	
rs1056503 ($P_{HWE} = 0.16$)	504	652			**0.004**
GG	281 (55.8%)	312 (47.9%)	1	1	
GT	197 (39.1%)	289 (44.3%)	**0.76 (0.59–0.97)**	**0.70 (0.53–0.93)**	
TT	26 (5.2%)	51 (7.8%)	**0.57 (0.34–0.93)**	0.65 (0.37–1.16)	
GT + TT	223 (44.2%)	340 (52.1%)	**0.73 (0.58–0.92)**	**0.70 (0.53–0.91)**	
rs9293337 ($P_{HWE} = 0.88$)	452	654			**0.007**
TT	251 (55.5%)	316 (48.3%)	1	1	
CT	173 (38.3%)	276 (42.2%)	0.79 (0.61–1.02)	**0.74 (0.55–0.99)**	
CC	21 (4.6%)	62 (9.5%)	**0.57 (0.35–0.92)**	**0.57 (0.31–0.98)**	
CT + CC	194 (42.9%)	338 (51.7%)	**0.75 (0.59–0.95)**	**0.71 (0.54–0.93)**	

Abbreviations: HWE, Hardy–Weinberg equilibrium; XRCC3, X-ray repair cross-complementing group 3; XRCC4, X-ray repair cross-complementing group 4.
*Adjusted by age, gender, education, family history of lung cancer, smoking status, environmental tobacco smoke, and mental and marital status.
Bold numerals: A P-value of <0.05 was considered statistically significant.

frequency of each gene polymorphism in the cases and controls, as well as the corresponding ORs for NSCLC. The variant genotypes of *XRCC3* rs861537 and *XRCC4* rs6869366 were found to be associated with an increased risk of NSCLC, with adjusted ORs of 1.46 (95% (CI): 1.10–1.94) and 1.86 (95% (CI): 1.18–2.91), respectively. Minor allele carriers of *XRCC3* rs1799794 and *XRCC4* rs1056503 and rs9293337 were inversely associated with NSCLC risk, with adjusted ORs of 0.73 (95% (CI): 0.55–0.98), 0.70 (95% (CI): 0.53–0.91) and 0.71 (95% (CI): 0.54–0.93), respectively.

Haplotypes and diplotypes of *XRCC3* and *XRCC4* SNPs and their association with NSCLC are shown in Table 3. In brief, the common haplotype served as the reference group, which was used to generate each OR. The distributions of the *XRCC3* (rs861539–rs1799796–rs861537–rs1799794) haplotype and *XRCC4* (rs3734091–rs1056503) haplotype were different between NSCLC cases and controls using a permutation test ($P<0.05$). For *XRCC3*, diplotypes of CGAA/CGAA, CAGA/CGAA and CAGG/TAAA were associated with the risk for NSCLC, with ORs of 1.93 (95% CI: 1.21–3.07), 1.82 (95% CI: 1.13–2.96) and 2.07 (95% CI: 1.05–4.08),

Table 3 Distributions of *XRCC3* and *XRCC4* diplotypes and haplotypes and their associations with non-small-cell lung cancer in Han population

	Frequency (%)			*Frequency (%)*		*Odds ratios*	
Haplotype	*Case*	*Control*	*Diplotype*	*Case*	*Control*	*95% confidence intervals*[a]	*P-value*
XRCC3 *(rs861539–rs1799796–rs861537–rs1799794)*($P_{permutation\ test} = 0.001$)							
CAGG	38.6	45.3	CAGG/CAGG	20.3	27.3	1	
CGAA	36.7	31.3	CAGG/CGAA	30.2	31.5	1.21 (0.84–1.74)	0.311
CAGA	14.5	12.2	CGAA/CGAA	15.4	10.4	**1.93 (1.21–3.07)**	**0.006**
TAAA	5.5	4.7	CAGA/CAGG	11.0	12.1	1.33 (0.82–2.15)	0.245
Others[b]	4.7	6.5	CAGA/CGAA	13.6	10.0	**1.82 (1.13–2.96)**	**0.015**
			CAGG/TAAA	5.7	3.5	**2.07 (1.05–4.08)**	**0.036**
			CAGA/TAAA	1.4	1.1	3.11 (0.93–10.35)	0.065
			Others[c]	2.4	4.1	0.68 (0.30–1.54)	0.352
XRCC4 *(rs3734091–rs1056503)*($P_{permutation\ test} = 0.013$)							
CG	70.8	66.8	CG/CG	49.1	44.1	1	
CT	22.9	27.9	CG/CT	34.5	39.4	0.76 (0.56–1.02)	0.067
AG	4.5	3.3	CG/AG	6.9	4.6	1.39 (0.76–2.52)	0.285
AT	1.8	2.0	CT/CT	4.2	5.7	0.67 (0.35–1.27)	0.220
			CG/AT	4.1	4.2	0.63 (0.33–1.21)	0.164
			Others	1.2	2.0	1.10 (0.35–3.46)	0.876

Abbreviations: XRCC3, X-ray repair cross-complementing group 3; XRCC4, X-ray repair cross-complementing group 4.
Bold numerals: A P-value of <0.05 was considered statistically significant.
[a]Adjusted by age, gender, education, family history of lung cancer, smoking status, environmental tobacco smoke, and mental and marital status.
[b]Frequency $<5\%$.
[c]Frequency $<2\%$.

respectively, as compared with diplotype CAGG/CAGG. For *XRCC4*, rs9293337, rs1056503, rs6869366 and rs3734091 were in linkage disequilibrium, and thus we chose rs3734091 and rs1056503 (linkage disequilibrium analysis showed $P<0.001$) for the diplotype analysis. However, the results did not show any association between these diplotypes and NSCLC risk compared with the CG/CG diplotype.

Gene–environment interactions

Gene–environment interactions were evaluated using the Clementine association rule mining analysis, in which NSCLC was a consequent and the diplotypes, age, gender, education levels, family history of lung cancer, ETS, marital status, mental status, smoking pack-years and the age of the first tobacco smoke were antecedents. The following parameters were used: minimum support = 5%; minimum confidence = 60%; and maximum number of antecedents = 4. As a result, we obtained 387 rules and chose the rules with a lift $>1.9\%$. The data are shown in Table 4. For the lift value, we could interpret the importance of numerous rules and further assessed the important variables with cancer. The data showed that male patients who smoked before the age of 20 years and had a history of over 26 pack-years, a family history of lung cancer or ETS exposure, CGAA/CGAA in *XRCC3* and CG/CG in *XRCC4* were associated with NSCLC risk. From all of these rules, the quantity of tobacco smoke was an essential component in determining the risk for developing NSCLC. Furthermore, logistic multifactor analysis showed that tobacco smokers over 26 pack-years with a family history of lung cancer history or ETS exposure, CGAA/CGAA in *XRCC3* and negative mental status were at risk of developing NSCLC.

DISCUSSION

In this case–control study of Chinese NSCLC patients and healthy control patients, we determined an association between *XRCC3* and *XRCC4* SNPs or other known risk factors and NSCLC susceptibility. We found five SNPs, including rs861537 and rs1799794 in *XRCC3* and rs6869366, rs1056503 and rs9293337 in *XRCC4*, to be associated with NSCLC risk in both adjusted and unadjusted models. For example, the rs1799794 G allele in the *XRCC3* gene was inversely associated with NSCLC risk (GG vs homozygote AA), whereas the rs861537 AG or AA genotype and *XRCC4* rs6869366 had a significantly increased NSCLC risk. In contrast, minor allele carriers of *XRCC4* rs1056503 and rs9293337 were inversely associated with NSCLC risk. In addition, 26 pack-year tobacco smokers, a family history of lung cancer, ETS exposure, CGAA/CGAA in *XRCC3* and a negative mental status were risk factors in NSCLC development. The data from this study indicate that gene–environment interactions have an important role in NSCLC development.

Previous studies showed that rs1799794 was a prognostic indicator for radiation and chemotherapy in NSCLC and a miR-328 binding site.[26,27] Moreover, the SNP marker rs861537 was in the subset of tagging SNPs identified by the Haploview Program and mapped to the intron region of *XRCC3*. This polymorphism was first shown to contribute to NSCLC risk in this study. However, to date, there is a lack of biological mechanistic and epidemiological data to support this finding. A recent study demonstrated that the rs861537 SNP was associated with the risk of lung, colorectal and breast cancer.[28] Another study showed that rs861539 was associated with G2 chromosomal radiosensitivity and could have a protective role in cancer susceptibility.[29] However, the same group 4 years later failed to repeat their previous data on the association between rs861539 and cancer risk or G(2) chromosomal radiosensitivity.[30] Meta-analysis[19,31] studies concluded that this SNP might have not been associated with lung cancer risk. However, our current study demonstrated that *XRCC3* rs861537 was associated with an increased risk of NSCLC, whereas the minor allele carriers of *XRCC3* rs1799794 were inversely associated with NSCLC risk.

Table 4 Association of gene–environment factors with NSCLC risk in Han population

Rule ID	Antecedent	Instances	Support (%)	Confidence (%)	Rule support (%)	Lift	Deployability
11	Family history of lung cancer plus 26 packs per year tobacco smoke	66	5.65	84.85	4.79	1.96	0.86
68	Family history of lung cancer plus 26 packs per year male tobacco smoker	66	5.65	84.85	4.79	1.96	0.86
293	rs3734091–rs1056503_CG/CG plus 26 packs per year smoker with education level of middle school or below or environmental tobacco smoke	87	7.44	83.91	6.24	1.93	1.20
264	Age starting smoke at <20 years old for 26 packs per year with rs3734091–rs1056503_CG/CG or environmental tobacco smoke	131	11.21	83.21	9.32	1.92	1.88
258	Age starting smoke at <20 years old for 26 packs per year with education level of middle school or below or environmental tobacco smoke	93	7.96	82.80	6.59	1.91	1.37
268	Smoker who married and age starting smoke at <20 years old for 26 packs per year or environmental tobacco smoke	185	15.83	82.70	13.09	1.91	2.74
117	rs861539–rs1799796–rs861537–rs1799794_CGAA/CGAA plus 26 packs per year smoker or environmental tobacco smoke	127	10.86	82.68	8.98	1.91	1.88
294	rs861539–rs1799796–rs861537–rs1799794_CGAA/CGAA plus 26 packs per year male smoker or environmental tobacco smoke	127	10.86	82.68	8.98	1.91	1.88

Abbreviation: NSCLC, non-small-cell lung cancer.

In this study, we did not find any association of rs1799796 with NSCLC risk. A previous study also suggested that there were no *XRCC3* polymorphisms associated with the susceptibility of urothelial bladder cancer.[32] However, this SNP has been associated with a reduced breast cancer risk.[20] Jacobsen *et al.*[18] found that rs1799796 combining with other two polymorphisms rs1799794 and rs861539 as a haplotype AAC that are associated with relatively high risk of lung cancer.

Our current study showed an association of rs1056503 with a reduced NSCLC risk. Liu *et al.*[33] did not find that this individual genotype was associated with glioma risk. In contrast, a three-locus interaction model showed that *LIG4* SNP rs1805388 (C>T), *XRCC4* SNP rs7734849 (A>T) and SNP rs1056503 (G>T) contributed to glioma susceptibility.[33] A functional analysis of *XRCC4* rs1056503 demonstrated that this polymorphism might have played a role in alternative splicing of mRNA.[34] This study suggested that the *XRCC4* rs9293337 genotype was associated with NSCLC risk and could be a novel marker for prevention and anticancer intervention studies. Further study is needed to explore the clinical significance of rs9293337 in NSCLC patients.

In addition, consistent with the data in other population studies, patients who smoked and carried the rs6869366 G allele had an increased risk for NSCLC ($P=0.004$). This association was not evident in non-smokers ($P=0.934$). This finding was similar to a Taiwanese study,[17] which indicated that this polymorphism may affect NSCLC risk after tobacco smoking. In this study, rs3734091 was not associated with NSCLC risk, even though some previous studies have assessed this SNP for its association with NSCLC risk.[16,17,35]

Although the association between a family history of cancer and lung cancer susceptibility has been previously reported widely,[36] our current study further confirmed such data. Moreover, we also found an association of NSCLC risk with other clinical and epidemiological data. Indeed, in the 1960s, one study showed that a repressed expression of emotions in cancer patients contributed to a type C personality ('cancer-prone'). In a recent prospective study,[37] patients with breast cancer tended to have an increased risk for bearing the 'high commitment' characteristic, which could contribute to cancer risk through immune and hormonal pathways. In this study, we found that negative personality was likely associated with NSCLC. We also confirmed that tobacco smoke (duration, smoking starting at young age and pack-years) was associated with NSCLC risk.

In this study, we analyzed our data by using the association rule, which is a classical algorithm of data mining. It has a strong ability to deal with incomplete data to discover patterns that are unknown and novel to investigators by providing a reference to understand and analyze the data. Using the association rule algorithm, we reduced the influence of missing values to find latent influencing factors. We then combined and analyzed the important variables using data mining software. After choosing variables, we focused on fewer indicators to form a model, allowing the data to be more consistent with logistic regression analyses.

In this study, we found that the risk factors in developing NSCLC were tobacco smokers over 26 pack-years with a family history of lung cancer, ETS exposure, CGAA/CGAA in *XRCC3* and negative mental status. Then, by observing the interaction results between diplotype *XRCC3* patients and environment factors (including smoking status, family history of lung cancer, ETS and mental status), we found that more than 26 pack-year tobacco smokers resulted in 7.08 times (after adjusting, $P<0.0001$) increased risk for developing NSCLC. With CGAA/CGAA in *XRCC3*, the risk was 24.71 times (after adjusting, $P=0.214$). We found that compared with the health controls, the OR value in the risk of suffering from NSCLC was 2.05 in the individuals with CGAA/CGAA in *XRCC3* and a family history of lung cancer, which was higher than 1.65 in those with a family history of lung cancer only. That value in the risk of suffering from NSCLC was 14.42 in the individuals with CGAA/CGAA in *XRCC3* and an exposure history of ETS, comparing with the health controls. It was higher than 4.83 in those with exposure history of ETS only. Compared with the health controls, the OR value in the risk of suffering from NSCLC was 2.18 in the individuals with CGAA/CGAA in *XRCC3* and negative mental status, which was higher than 1.73 in those with negative mental status only. Even though the *P*-value was >0.05, without considering the 95% CI of OR value, it indicated that the diplotype of *XRCC3* CGAA/CGAA had a synergistic effect with smoke status, family history of lung cancer, ETS and mental status.

This study was also subject to several methodological limitations. For example, given the retrospective nature of this study design, smoking behaviors were recalled by participants and were subject to information bias. In addition, unequal recall among cases and controls could have led to an overestimation of the observed associations. The personality status data, which was collected from

the patients' self-assessment in this study, were not specifically collected using the examining personality characteristics, and thus their association with NSCLC risk might have been aggrandized. Further study is needed to examine this association using a measuring scale.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported in part by grants from the Natural Science Foundation of China (Nos 30771845 and 81172766), the US NIH National Institute of Environmental Health Sciences (No. ES011667), National Cancer Institute (Nos CA90833 and CA09142) and the Alper Research Center for Environmental Genomics. We thank all the medical staff in the Department of Thoracic and Cardiac Surgery, The First Clinical Medical College, the Affiliated Union Hospital of Fujian Medical University and Fuzhou General Hospital for collecting epidemiological data and biological specimens for the study.

1 Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. & Forman, D. Global cancer statistics. *CA Cancer J. Clin.* **61,** 69–90 (2011).
2 Tokuhata, G. K. & Lilienfeld, A. M. Familial aggregation of lung cancer in humans. *J. Natl Cancer Inst.* **30,** 289–312 (1963).
3 La Maestra, S., Kisby, G. E., Micale, R. T., Johnson, J., Kow, Y. W., Bao, G. *et al.* Cigarette smoke induces DNA damage and alters base-excision repair and tau levels in the brain of neonatal mice. *Toxicol. Sci.* **123,** 471–479 (2011).
4 Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.* **27,** 247–254 (2001).
5 Ferguson, D. O. & Alt, F. W. DNA double strand break repair and chromosomal translocation: lessons from animal models. *Oncogene* **20,** 5572–5579 (2001).
6 Cui, X., Brenneman, M., Meyne, J., Oshimura, M., Goodwin, E. H. & Chen, D. J. The XRCC2 and XRCC3 repair genes are required for chromosome stability in mammalian cells. *Mutat. Res.* **434,** 75–88 (1999).
7 Kurumizaka, H., Ikawa, S., Nakada, M., Eda, K., Kagawa, W., Takata, M. *et al.* Homologous-pairing activity of the human DNA-repair proteins Xrcc3.Rad51C. *Proc. Natl Acad. Sci. USA* **98,** 5538–5543 (2001).
8 Tebbs, R. S., Zhao, Y., Tucker, J. D., Scheerer, J. B., Siciliano, M. J., Hwang, M. *et al.* Correction of chromosomal instability and sensitivity to diverse mutagens by a cloned cDNA of the XRCC3 DNA repair gene. *Proc. Natl Acad. Sci. USA* **92,** 6354–6358 (1995).
9 Brenneman, M. A., Weiss, A. E., Nickoloff, J. A. & Chen, D. J. XRCC3 is required for efficient repair of chromosome breaks by homologous recombination. *Mutat. Res. DNA Rep.* **459,** 89–97 (2000).
10 Kawabata, M., Kawabata, T. & Nishibori, M. Role of recA/RAD51 family proteins in mammals. *Acta. Med. Okayama* **59,** 1–9 (2005).
11 Mari, P. O., Florea, B. I., Persengiev, S. P., Verkaik, N. S., Brüggenwirth, H. T., Modesti, M. *et al.* Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. *Proc. Natl Acad. Sci. USA* **103,** 18597–18602 (2006).
12 Frank, K. M., Sekiguchi, J. M., Seidl, K. J., Swat, W., Rathbun, G. A., Cheng, H. L. *et al.* Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396,** 173–177 (1998).
13 Yin, Q. H., Liu, C., Li, L., Zu, X. Y. & Wang, Y. J. Association between the XRCC3 T241M polymorphism and head and neck cancer susceptibility: a meta-analysis of case–control studies. *Asian. Pac. J. Cancer Prev.* **13,** 5201–5205 (2012).
14 Romanowicz-Makowska, H., Brys, M., Forma, E., Maciejczyk, R., Polac, I., Samulak, D. *et al.* Single nucleotide polymorphism (SNP) Thr241Met in the XRCC3 gene and breast cancer risk in Polish women. *Pol. J. Pathol.* **63,** 121–125 (2012).
15 Mittal, R. D., Gangwar, R., Mandal, R. K., Srivastava, P. & Ahirwar, D. K. Gene variants of XRCC4 and XRCC3 and their association with risk for urothelial bladder cancer. *Mol. Biol. Rep.* **39,** 1667–1675 (2012).
16 Ryk, C., Kumar, R., Thirumaran, R. K. & Hou, S. M. Polymorphisms in the DNA repair genes XRCC1, APEX1, XRCC3 and NBS1, and the risk for lung cancer in never- and ever-smokers. *Lung. Cancer* **54,** 285–292 (2006).
17 Hsu, N. Y., Wang, H. C., Wang, C. H., Chang, C. L., Chiu, C. F., Lee, H. Z. *et al.* Lung cancer susceptibility and genetic polymorphism of DNA repair gene XRCC4 in Taiwan. *Cancer Biomark.* **5,** 159–165 (2009).
18 Jacobsen, N. R., Raaschou-Nielsen, O., Nexø, B., Wallin, H., Overvad, K., Tjønneland, A. *et al.* XRCC3 polymorphisms and risk of lung cancer. *Cancer Lett.* **15,** 67–72 (2004).
19 Shi, C. L., Li, R., Xiong, L. W., Gu, A. Q., Han, B. H. & Gu, W. Lack of association between XRCC3 rs861539 (C>T) polymorphism and lung cancer risks: an update meta-analysis. *Tumour Biol.* **34,** 1819–1824 (2013).
20 He, X. F., Wei, W., Su, J., Yang, Z. X., Liu, Y., Zhang, Y. *et al.* Association between the XRCC3 polymorphisms and breast cancer risk: meta-analysis based on case–control studies. *Mol. Biol. Rep.* **39,** 5125–5134 (2012).
21 Improta, G., Sgambato, A., Bianchino, G., Zupa, A., Grieco, V., La Torre, G. *et al.* Polymorphisms of the DNA repair genes XRCC1 and XRCC3 and risk of lung and colorectal cancer: a case–control study in a Southern Italian population. *Anticancer Res.* **28,** 2941–2946 (2008).
22 Qian, B., Zhang, H., Zhang, L., Zhou, X., Yu, H. & Chen, K. Association of genetic polymorphisms in DNA repair pathway genes with non-small cell lung cancer risk. *Lung Cancer* **73,** 138–146 (2011).
23 Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L. & Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 106–120 (2004).
24 de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., Altshuler, D. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37,** 1217–1223 (2005).
25 Peltonen, L. & McKusick, V. A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291,** 1224–1229 (2001).
26 Arora, S., Ranade, A. R., Tran, N. L., Nasser, S., Sridhar, S., Korn, R. L. *et al.* MicroRNA-328 is associated with (non-small) cell lung cancer (NSCLC) brain metastasis and mediates NSCLC migration. *Int. J. Cancer* **129,** 2621–2631 (2011).
27 Dacic, S., Kelly, L., Shuai, Y. & Nikiforova, M. N. miRNA expression profiling of lung adenocarcinomas: correlation with mutational status. *Mod. Pathol.* **23,** 1577–1582 (2010).
28 Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol.* **12,** 477–488 (2011).
29 Wilding, C. S., Curwen, G. B., Tawn, E. J., Sheng, X., Winther, J. F., Chakraborty, R. *et al.* Influence of polymorphisms at loci encoding DNA repair proteins on cancer susceptibility and G2 chromosomal radiosensitivity. *Environ. Mol. Mutagen.* **48,** 48–57 (2007).
30 Curwen, G. B., Murphy, S., Tawn, E. J., Winther, J. F. & Boice, J. D. Jr. A study of DNA damage recognition and repair gene polymorphisms in relation to cancer predisposition and G2 chromosomal radiosensitivity. *Environ. Mol. Mutagen.* **52,** 72–76 (2011).
31 Sun, H., Qiao, Y., Zhang, X., Xu, L., Jia, X., Sun, D. *et al.* XRCC3 Thr241Met polymorphism with lung cancer and bladder cancer: a meta-analysis. *Cancer Sci.* **101,** 1777–1782 (2010).
32 Mittal, R. D., Gangwar, R., Mandal, R. K., Srivastava, P. & Ahirwar, D. K. Gene variants of XRCC4 and XRCC3 and their association with risk for urothelial bladder cancer. *Mol. Biol. Rep.* **39,** 1667–1675 (2012).
33 Liu, Y., Zhou, K., Zhang, H., Shugart, Y. Y., Chen, L., Xu, Z. *et al.* Polymorphisms of LIG4 and XRCC4 involved in the NHEJ pathway interact to modify risk of glioma. *Hum. Mutat.* **29,** 381–389 (2008).
34 Lee, P. H. & Shatkay, H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* **36,** D820–D824 (2008).
35 Hsieh, Y. Y., Bau, D. T., Chang, C. C., Tsai, C. H., Chen, C. P. & Tsai, F. J. XRCC4 codon 247* A and XRCC4 promoter − 1394* T related genotypes but not XRCC4 intron 3 gene polymorphism are associated with higher susceptibility for endometriosis. *Mol. Reprod. Dev.* **75,** 946–951 (2008).
36 Matakidou, A., Eisen, T. & Houlston, R. S. Systematic review of the relationship between family history and lung cancer risk. *Br. J. Cancer.* **93,** 825–833 (2005).
37 Eskelinen, M. & Ollonen, P. Assessment of 'cancer-prone personality' characteristics in healthy study subjects and in patients with breast disease and breast cancer using the commitment questionnaire: a prospective case–control study in Finland. *Anticancer Res.* **31,** 4013–4017 (2011).