Tagging single nucleotide polymorphisms in excision repair cross-complementing group 1 (*ERCC1*) and risk of primary lung cancer in a Chinese population

Hongxia Ma[a,*], Liang Xu[b,*], Jing Yuan[d], Minhua Shao[c], Zhibin Hu[a], Feng Wang[d], Yi Wang[c], Wentao Yuan[b], Ji Qian[c], Ying Wang[b], Pengcheng Xun[a], Hongliang Liu[c], Weihong Chen[d], Lin Yang[b], Guangfu Jin[a], Xiang Huo[a], Feng Chen[a], Yin Yao Shugart[e], Li Jin[b,c], Qingyi Wei[f], Tangchun Wu[d], Hongbing Shen[a], Wei Huang[b] and Daru Lu[c]

Background and objective **Low nucleotide excision repair (NER) capacity has been associated with increased risk of lung cancer. Excision repair cross-complementing group 1 (ERCC1) is one of the NER core enzymes, and polymorphisms in *ERCC1* may lead to altered repair function of the enzyme and therefore confer predisposition to cancer. The goal of this study was to test the hypothesis that common variants in *ERCC1* were associated with lung cancer risk.**

Methods **The genotyping analyses for 7 selected single nucleotide polymorphisms in *ERCC1* using the TaqMan assay was conducted in a case–control study of 1010 patients with incident lung cancer and 1011 cancer-free controls in a Chinese population.**

Results **We found that the variant genotypes of the rs3212948 C allele were associated with significantly decreased risk of lung cancer [adjusted odds ratio (OR) = 0.73 (95% CI = 0.60–0.88) for CG; 0.96 (95% CI = 0.65–1.41) for CC and 0.76 (95% CI = 0.63–0.91) for CG/CC, compared with the GG genotype]. Similarly, a significant protective effect was also evident for the variant genotypes of rs1007616 C/T [adjusted OR = 0.72 (95% CI = 0.59–0.89) for CT; 0.90 (95% CI = 0.61–1.35) for TT and 0.75 (95% CI = 0.62–0.91) for CT/TT, compared with the CC genotype]. Stratified analysis revealed that the protective effects of these 2 single nucleotide polymorphisms were both more evident among young patients and patients without family history of cancer. Consistently, when assessing each unique haplotype compared with the most common haplotype 'TAGCACG', lung cancer risk was significantly decreased among patients who carried the haplotype 'TCCCATT' with the variant rs3212948C and rs1007616T alleles (*P* value = 0.0340, *P*-sim = 0.0325, adjusted OR = 0.78; 95% CI = 0.63–0.97).**

Conclusion **These findings indicate that *ERCC1* polymorphisms may contribute to the etiology of lung cancer. Further functional studies were warranted to elucidate the mechanism of the associations.**
Pharmacogenetics and Genomics **17:417–423**

Pharmacogenetics and Genomics 2007, **17**:417–423

Keywords: case–control study, DNA repair, genetic susceptibility, haplotype, molecular epidemiology

[a]Department of Epidemiology and Biostatistics, Cancer Research Center of Nanjing Medical University, Nanjing, [b]Department of Genetics, Chinese National Human Genome Center at Shanghai, [c]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, [d]Institute of Occupational Medicine, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [e]Epidemiology Department, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland and [f]Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas, USA

Correspondence to Professor Daru Lu, State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Handan Rd., Shanghai 200433, China
Tel/fax: +86 21 65642799; e-mail: drlu@fudan.edu.cn

*Ma H. and Xu L. contributed equally to this work.

Received 9 January 2006 **Accepted** 10 August 2006

Introduction

Cellular DNA is routinely damaged by endogenous or exogenous mutagens such as ultraviolet light, tobacco smoke, and reactive oxygen species [1,2]. DNA damage must be repaired to enable the cell to maintain genomic integrity [3]. Alternatively, unrepaired DNA damage can accumulate, resulting in apoptosis or leading to unregulated cell growth and carcinogenesis [2]. Complex pathways involving numerous enzymes to perform nucleotide excision repair (NER) is a major cellular defense mechanism against DNA damage from bulky DNA adducts induced by chemical carcinogens, such as polycyclic aromatic hydrocarbons from tobacco smoke [4]. Studies have shown that impaired DRC (DNA repair capacity) is associated with increased risk of smoking-related lung cancer [5,6].

1744-6872

Excision repair cross-complementation group 1 (ERCC1), a highly conserved enzyme and the major protein involved in NER, is the lead component of the NER process and acts in a complex with XPF (Xeroderma pigmentosum complementation group F) to make the incision at 5′ to the lesion site [7,8]. Therefore, a defect in ERCC1 may cause severe DNA repair deficiency [7]. Unlike other NER proteins, ERCC1 is also involved in other DNA repair pathways such as homologous recombination, double-strand break repair, and the repair of interstrand cross-links [7]. Cells from ERCC1-knockout mice are highly sensitive to DNA cross-linking agents and have profound cell cycle abnormalities and a reduced frequency of S-phase-dependent illegitimate chromosome exchange [9,10]. It has been shown that higher *ERCC1* mRNA levels are associated with a more active DNA repair process [11] and that low expression of *ERCC1* mRNA was shown to be associated with a higher risk of lung cancer [12].

Single nucleotide polymorphisms (SNPs) were found in nearly all human DNA repair genes, some of which had been shown to modulate levels of DNA damage, individual DRC, and cancer risk [2,13–15]. Although no common nonsynonymous SNP has been found in *ERCC1*, several published studies evaluated the associations between common SNPs in the *ERCC1* noncoding regions and cancer risk [16–19]. In a recently published case–control study of 1752 Caucasian lung cancer patients and 1358 controls, Zhou *et al.* [18] genotyped two SNPs of *ERCC1,* rs3212986 (3′UTR) and rs11615 (synonymous SNP, codon 118) and found no main effects of these two variants on lung cancer risk. Similarly, in a small case–control study of 122 lung cancer cases and 122 cancer-free controls in Xuan Wei, China, Shen *et al.* [19] also failed to find a significant association between the *ERCC1* rs3212961 (intron 5) and rs3212948 (intron 3) variants and lung cancer risk. A limited number of SNPs, however, were investigated in the former study and a small sample size was included in the later study. To further investigate the association between SNPs of *ERCC1* and risk of lung cancer for identifying potential genetic markers for lung cancer susceptibility, we conducted a case–control study of 1010 incident lung cancer cases and 1011 age and sex frequency-matched cancer-free controls in a Chinese population, and we genotyped six *ERCC1* SNPs selected from the Environmental Genome Project (EGP) database and also a newly identified tagging SNPs (tagSNPs) of the gene. We tested the hypothesis that genetic variants of *ERCC1* are associated with risk of developing lung cancer.

Materials and methods

Study population

All the participants included in this hospital-based case–control study were genetically unrelated ethnic Han Chinese who provided a written informed consent. Eligible patients had newly diagnosed incident lung cancer according to the National Diagnosis Standard for Lung Cancer, who were consecutively recruited between July 2002 and November 2004 from the Cancer Hospital of Jiangsu Province, the First Affiliated Hospital of Nanjing Medical University, the Shanghai Cancer Hospital, and the Wuhan Zhongnan Hospital, without the restrictions of age, sex, and histology. Those who had previous cancer, metastasized cancer, and previous radiotherapy or chemotherapy were excluded. A total of 1299 cases with histopathologically confirmed lung cancer were recruited, of whom 1010 patients consented to participate in the study and provided blood samples, resulting in a response rate of 77.8% (1010/1299). The 1011 cancer-free controls consisted of outpatients with diseases other than cancer in other departments of the same hospital (general surgery, gynecology, internal medicine, orthopedics, and otorhinolaryngology) during the same time period when the cases were recruited. These controls reported no cancer history and were frequency-matched to the cases on age ($\pm$ 5 years), sex, and residential area (urban or countryside), and the response rate was 81.3% (1011/1244).

Each participant was scheduled for an interview, and a structured questionnaire was administered by interviewers to collect information on demographic data and environmental exposure history including tobacco smoking. Those who had smoked less than one cigarette per day and shorter than 1 year in their lifetime were defined as nonsmokers, otherwise, they were considered as smokers. Those smokers who quit for more than 1 year were considered former smokers. Pack-years [(cigarettes per day $\div$ 20) $\times$ years smoked] were calculated to indicate the cumulative smoking dose and the smokers being further dichotomized by the cumulative dose of 29 pack-years according to the distribution of controls. Family history of cancer was defined as any reported cancer in first-degree relatives (parents, siblings, or children). After interview, approximately 5 ml venous blood sample was collected from each participant. The study was approved by the institutional review boards of Nanjing Medical University, Fudan University, and Tongji Medical College of Huazhong University of Science and Technology.

Single nucleotide polymorphisms selection in *ERCC1*

As many of the SNPs in the human genome cosegregate with other nearby SNPs, it is possible to use a small number of markers, that is, the tagSNPs, to capture the majority of the sequence variation in a gene. For the *ERCC1* gene (AF512555), we used the resequencing data of 90 individuals in the EGP database (*http://egp.gs.washington.edu/*) to select SNPs from the reported polymorphisms on the basis of calculation of pairwise linkage disequilibrium (LD). Totally 30 common SNPs are present with a minor allele frequency (MAF) > 0.1 in

the database and we used a greedy algorithm proposed by Carlson *et al.* [20] to choose the tagSNPs. By using this algorithm, we selected an informative set of common (i.e. MAF > 0.1) tagSNPs, and these LD-based tagSNPs were considered more representative than an equivalent number of either haplotype-selected htSNPs or randomly selected SNPs for association studies [20]. As a result, we selected six tagSNPs in the *ERCC1* gene with a minimal LD parameter r^2 threshold of 0.5. In addition, we included an additional one newly identified variant from our previous SNPs screening project in the Chinese population (*ERCC1-17172*). Therefore, a total of seven polymorphisms (Table 1) were genotyped for the 1010 lung cancer patients and 1011 cancer-free controls included in the final analysis.

Laboratory assays

Genotyping was performed by the 5′-nuclease (TaqMan) assay, using the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, California, USA), in 384-well format, in Chinese National Human Genome Center at Shanghai, China. The TaqMan primers and probes were designed using the Primer Express Oligo Design software v2.0 (ABI PRISM). Primer and probe sequences are available upon request. PCR reactions were carried out in a reaction volume of 5 µl containing 5 ng DNA, 2.5 µl 2× TaqMan Universal PCR Master Mix (Applied Biosystems), 0.083 µl 40× Assay Mix. PCR reaction conditions included 95°C for 10 min followed by 20 cycles of 15 s at 92°C and 1 min at 60°C followed by 30 cycles of 15 s at 89°C and 1.5 min at 60°C. Two blank (water) controls and two duplicated samples in each 384-well format were used for quality control. The intensity of each SNP should meet the criteria of three clear clusters in two scales generated by SDS software (ABI). The genotypes were successful in $> 95\%$ participants for six SNPs and only one SNP (rs1007616) were lower than 90% (Table 1).

Statistical analyses

Differences in selected demographic variables, smoking status, pack-years of smoking, family history of cancer, and frequencies of the *ERCC1* genotypes, alleles, and haplotypes between the cases and controls were evaluated by using the χ^2 test. For each polymorphism, deviation from genotype frequency distribution in controls from those expected under Hardy–Weinberg equilibrium was also assessed using the χ^2 test. The associations between *ERCC1* variants and lung cancer risk were estimated by computing the odds ratios (ORs) and 95% confidence intervals (CIs) from both univariate and multivariate logistic regression analyses with adjustment for age, sex, pack-years of smoking, and family history of cancer. All of the statistical analyses were performed with Statistical Analysis System software (v.8.0e; SAS Institute, Cary, North Carolina, USA). Haplotype analyses were conducted using Haplo.stats (*http://www.mayo.edu/stagen*), which is a score test based on generalized linear models (GLMs) that test for associations between haplotypes and disease of interest under the null hypothesis of no haplotype effect without any assumption about mode of inheritance. This software also provides several different global and haplotype-specific tests for association and allows the possibility to include several nongenetic covariates.

Results

The characteristics of the 1010 lung cancer patients and 1011 cancer-free controls have been described elsewhere [21]. Briefly, the lung cancer cases and controls appeared to be adequately matched on age and sex ($P = 0.98$ and 0.30, respectively). More smokers, however, were present among the cases (68.8%) than were among the controls (52.2%), and more cases (44.5%) smoked greater than 30 pack-years than did the controls (25.4%), and this difference was statistically significant ($P < 0.0001$). Furthermore, lung cancer cases were significantly more likely than the controls to report a family history of cancer (17.1% versus 12.8%; $P = 0.006$) in their first-degree relatives, and this difference accounted for a significantly 41% increased lung cancer risk (OR = 1.41, 95% CI = 1.10–1.81). Among the lung cancer cases, 430 (42.6%) were adenocarcinoma, 335 (33.2%) were squamous cell carcinoma, 65 (6.4%) were small-cell carcinoma, and 180 (17.8%) were other histopathologic types, including large cell, mixed cell, or undifferentiated carcinomas.

Table 1 Primary information of genotyped SNPs of ERCC1

Gene (accession no.) and locus	NCBI rs no.	Location	Base change	MAF of cases	MAF of controls	MAF of EGP[a]	*P* for HWE	Genotyped (%)
ERCC1	rs3212930	001454, close to 5′UTR	T>C	0.094	0.092	0.20	0.789	95.9
(AF512555)	rs2298881	002148, 5′UTR	C>A	0.399	0.391	0.20	0.190	98.0
19q13.2–q13.3	rs3212948	004703, intron 2	G>C	0.231	0.267	0.38	0.171	97.9
	rs3212951	005027, intron 2	C>T	0.183	0.188	0.06	0.799	98.6
	rs3212955	005569, intron 3	A>G	0.316	0.295	0.29	0.647	97.6
	rs1007616	015972, intron 8	C>T	0.223	0.262	0.42	0.812	86.2
	NA[b]	017172, close to 3′UTR	T>G	0.440	0.426	–	0.597	98.3

HWE, Hardy–Weinberg equilibrium; EGP, Environmental Genome Project; ERCC1, excision repair cross-complementing group 1; MAF, minor allele frequencies; NCBI, National Center for Biotechnology Information; NA, not available; SNP, single nucleotide polymorphism; UTR, untranslated region.
[a]EGP, US Environmental Genome Project SNP database (see http://egp.gs.washington.edu/directory.html).
[b]NA: Not available in NCBI as well as in EGP database because this SNP was newly identified by the authors in China.

Table 2 Logistic regression analysis of associations between the genotypes in selected SNPs and lung cancer risk

Genotype	Cases no. (%)	Controls no. (%)	*P*	Adjusted OR (95% CI)[a]
rs3212930	963	976		
TT	787 (81.7)	805 (82.5)	0.635	1.00
CT	170 (17.6)	162 (16.6)		1.05 (0.82–1.34)
CC	6 (0.6)	9 (0.9)		0.73 (0.25–2.13)
CT/CC	176 (18.2)	171 (17.5)	0.795	1.03 (0.81–1.31)
rs2298881	987	994		
CC	349 (35.4)	359 (36.1)	0.835	1.00
AC	488 (49.4)	493 (49.6)		1.00 (0.82–1.43)
AA	150 (15.2)	142 (14.3)		1.08 (0.81–1.43)
AC/AA	638 (64.6)	635 (63.9)	0.867	1.02 (0.84–1.23)
rs3212948	992	986		
GG	594 (59.9)	521 (52.8)	0.005	1.00
CG	338 (34.1)	403 (40.9)		0.73 (0.60–0.88)
CC	60 (6.1)	62 (6.3)		0.96 (0.65–1.41)
CG + CC	398 (40.2)	465 (47.2)	0.003	0.76 (0.63–0.91)
rs3212951	994	998		
CC	667 (67.1)	657 (65.8)	0.712	1.00
CT	290 (29.2)	307 (30.8)		0.92 (0.75–1.12)
TT	37 (3.7)	34 (3.4)		1.04 (0.63–1.70)
CT + TT	327 (32.9)	341 (34.2)	0.455	0.93 (0.77–1.13)
rs3212955	984	988		
AA	462 (47.0)	494 (50.0)	0.370	1.00
AG	423 (43.0)	405 (41.0)		1.09 (0.90–1.32)
GG	99 (10.0)	89 (9.0)		1.20 (0.87–1.65)
AG + GG	38 (53.0)	494 (50.0)	0.268	1.11 (0.92–1.33)
rs1007616	835	908		
CC	513 (61.4)	493 (54.3)	0.009	1.00
CT	270 (32.3)	354 (39.0)		0.72 (0.59–0.89)
TT	52 (6.2)	61 (6.7)		0.90 (0.61–1.35)
CT/TT	322 (38.5)	415 (45.7)	0.004	0.75 (0.62–0.91)
ERCC1-17172	991	996		
TT	301 (30.4)	332 (33.3)	0.304	1.00
TG	508 (51.3)	479 (48.1)		1.16 (0.94–1.42)
GG	182 (18.4)	185 (18.6)		1.08 (0.83–1.40)
TG/GG	690 (69.7)	664 (66.7)	0.205	1.13 (0.93–1.38)

CI, confidence interval; SNP, single nucleotide polymorphism; OR, odds ratio.
[a]Adjusted for age, sex, pack-years of smoking, and family history of cancer.

As shown in Table 1, all *ERCC1* genotype distributions in the controls were consistent with those expected from the Hardy–Weinberg equilibrium model. Almost all SNPs in this study population had a MAF 10% greater or lesser than those reported in the EGP SNP database (*http://egp.gs.washington.edu/*), which may reflect either ethnic differences or frequency bias owing to small sample sizes from which the database derived.

The allele frequencies and genotype distributions of *ERCC1* polymorphisms in cases and controls are summarized in Table 2. In the single locus analyses, only the distributions of two SNPs were significantly different between the cases and the controls ($P = 0.005$ for rs3212948 and $P = 0.009$ for rs1007616). Multivariate logistic regression analyses revealed that significantly protective effects were associated with the variant genotypes of rs3212948G/C [adjusted OR = 0.73 (95% CI = 0.60–0.88) for CG; 0.96 (95% CI = 0.65–1.41) for CC and 0.76 (95% CI = 0.63–0.91) for CG/CC genotypes, compared with GG genotype], and with the variant genotypes of rs1007616C/T [adjusted OR = 0.72 (95% CI = 0.59–0.89) for CT; 0.90 (95% CI = 0.61–1.35) for TT and 0.75 (95% CI = 0.62–0.91) for CT/TT genotypes, compared with CC genotype] (Table 2).

We further evaluated the associations of the rs3212948CG/CC and rs1007616CT/TT variant genotypes with lung cancer risk stratified by selected variables and histological types. As shown in Table 3, the effect of combined variant genotypes were more evident in both

Table 3 Stratified analyses between the *ERCC1* rs3212948 and rs1007616 genotypes and lung cancer risk

Variables	ERCC1 rs3212948				ERCC1 rs1007616			
	Cases/control (992/986)		Adjusted OR (95% CI)[a]		Cases/controls (835/908)		Adjusted OR (95% CI)[a]	
	GG	CG/CC	GG	GG/CC	CC	CT/TT	CC	CT/TT
	n	*n*			*n*	*n*		
Age (years)								
≤ 60	298/253	190/234	1.00	0.68 (0.52–0.88)	255/234	150/200	1.00	0.68 (0.51–0.90)
>60	296/268	208/231	1.00	0.85 (0.65–1.10)	258/259	172/215	1.00	0.84 (0.64–1.11)
Sex								
Male	462/398	304/339	1.00	0.77 (0.63–0.96)	396/360	250/295	1.00	0.77 (0.61–0.96)
Female	132/123	94/126	1.00	0.69 (0.48–0.99)	117/133	72/120	1.00	0.69 (0.47–1.02)
Pack-years of smoking								
0	180/245	126/224	1.00	0.75 (0.56–1.01)	162/244	94/204	1.00	0.69 (0.50–0.94)
1–29	152/142	90/126	1.00	0.65 (0.45–0.93)	132/125	86/105	1.00	0.79 (0.54–1.16)
30+	262/134	182/115	1.00	0.82 (0.60–1.12)	219/124	142/106	1.00	0.75 (0.54–1.06)
Family history of cancer								
No	492/447	328/413	1.00	0.72 (0.59–0.88)	419/425	261/370	1.00	0.72 (0.58–0.89)
Yes	102/74	70/52	1.00	0.99 (0.61–1.62)	94/68	61/45	1.00	0.95 (0.56–1.61)
Histological types								
Adenocarcinomas	246/521	176/465	1.00	0.79 (0.63–1.01)	223/493	144/415	1.00	0.77 (0.60–0.99)
Squamous cell	196/521	134/465	1.00	0.80 (0.61–1.41)	168/493	110/415	1.00	0.81 (0.61–1.08)
Small cell	42/521	22/465	1.00	0.62 (0.36–1.06)	36/493	23/415	1.00	0.79 (0.45–1.36)
Other carcinomas	110/521	66/465	1.00	0.67 (0.48–0.94)	86/493	45/415	1.00	0.62 (0.42–0.92)

CI, confidence interval; ERCC1, excision repair cross-complementing group 1; OR, odds ratio.
[a]Adjusted for age, sex, pack-years of smoking, and family history of cancer.

Table 4 Associations between risk of lung cancer and frequencies of inferred haplotypes on the basis of the observed genotypes in lung cancer cases and cancer-free controls

[a]Haplotype	All case patients		Controls		*P*	*P*-sim[b]	Adjusted OR (95% CI)[c]
	No.	Frequencies	No.	Frequencies			
TAGCACG	197	0.26097	200	0.23727	0.1409	0.1420	1.00
TCCCATT	153	0.20283	199	0.23599	0.0340	0.0325	0.78 (0.63–0.97)
TCGCGTT	10	0.01368	9	0.01051	0.4050	0.4039	1.27 (0.64–2.53)
TAGTACG	90	0.11897	105	0.12383	0.6952	0.6967	0.91 (0.71–1.16)
CCGCGCT	64	0.08438	71	0.08337	0.9855	0.9868	0.91 (0.68–1.20)
TCGTACT	25	0.03309	30	0.03587	0.5281	0.5331	0.82 (0.54–1.24)
TCCCACG	19	0.02548	21	0.02484	0.7954	0.7980	0.97 (0.60–1.55)
TCGTACG	18	0.02443	16	0.01843	0.1584	0.1601	1.22 (0.71–2.09)
TCGCGCT	147	0.19418	155	0.18384	0.4082	0.4190	0.97 (0.78–1.21)

CI, confidence interval; OR, odds ratio.
[a]Polymorphic bases were in 5′–3′ order as listed in Table 2.
[b]Simulation-based *P*-value.
[c]Adjusted for age, sex, pack-years of smoking, and family history of cancer.

young individuals [adjusted OR = 0.68, 95% (CI = 0.52–0.88) for rs3212948CG/CC and adjusted OR = 0.68 (95% CI = 0.51–0.90) for rs1007616CT/TT], and individuals without a family history of cancer [adjusted OR = 0.72 (95% CI = 0.59–0.88) for rs3212948CG/CC and adjusted OR = 0.72 (95% CI = 0.58–0.89) for rs1007616CT/TT].

Haplotypes with a frequency less than 0.01 were omitted, and the remaining nine haplotypes were analyzed with Haplo.stats. After adjusting for age, sex, pack-years of smoking, and family history of cancer, the risk of lung cancer was significantly decreased among individuals carrying the haplotype 'TCCCATT', with the variant rs3212948C and rs1007616T alleles (*P*-value = 0.034, *P*-sim = 0.033, adjusted OR = 0.78; 95% CI = 0.63–0.97), compared with those carrying the most common haplotype 'TAGCACG' (Table 4). Furthermore, global score tests were also performed using Haplo.stats. When no covariates were included, the global haplo score was 9.47 (d.f. = 9) with an asymptotic *P*-value of 0.395 and a simulation-based *P*-value of 0.459 (10 000 permutations). When adjusted for age, sex, pack-years of smoking, and family cancer history, the global haplo score was 9.60 with an asymptotic *P*-value of 0.384 and a stimulation-based *P*-value of 0.387 (10 000 permutations).

Discussion

In this relatively large-scale case–control study, we found, for the first time, that the rs3212948 (G/C) and rs1007616 (C/T) polymorphisms of *ERCC1* gene were significantly associated with a decreased risk of lung cancer in a Chinese population. Stratified analysis revealed that the protective effects of these two SNPs were both more evident among young individuals and those without a family history of cancer. Furthermore, using the haplotype-based approach, we found that *ERCC1* haplotype 'TCCCATT' was a protective predictor for risk of lung cancer. These findings indicate that SNPs of the *ERCC1* gene may be biomarkers for susceptibility to lung cancer, which warrants further validation with functional studies.

Human lung cancer is a well-known example of environmentally induced carcinogenesis, such as by environmental pollutants and tobacco smoke [1]. Tobacco carcinogens induce various types of DNA damage, including DNA adducts, strand breaks, cross-links, and recombination, which stimulate different DNA repair pathways to remove the damage and maintain the genomic stability [18]. As a major protein involved in NER, ERCC1 forms a heterodimer with XPF to execute the incision into the strand at 5′ of the site of damage, essential to the accurate DNA repair [22]. Clinically, lower ERCC1 levels are associated with improved disease response and improved relapse-free survival [23–26], which suggest that levels of ERCC1 may correlate with DNA repair activity, indicating an important role of ERCC1 in the development of cancer.

Several studies have investigated the associations between *ERCC1* polymorphisms and human cancer risk, but mostly focusing on only one or two variants in the coding region or the 3′ untranslated region of *ERCC1* gene. For example, an A to C polymorphism at nucleotide 8092 (rs3212986) in the 3′ untranslated region is thought to affect ERCC1 mRNA stability and found to be associated with an advanced age onset of adult glioma [16], but it does not appear to play an important role in the etiology of head and neck cancers [17] and lung cancer [18]. As no common nonsynonymous SNP was found in the *ERCC1* coding region, we used a haplotype-based analysis, which does not require a priori identification of functional SNPs [27], to assess the association between genetic variation in the *ERCC1* gene and risk of lung cancer. Haplotype-based analysis assumes that genotypes of unassayed, risk-related SNPs may be linked with one or more assayed SNPs. Kaplan and Morris [28] showed that haplotype-based tests could have a greater power than unphased tests in the case when the disease locus has multiple

disease-causing alleles. In such an analysis, statistical power to detect unassayed, disease-associated polymorphisms depends on the correlation (r^2) between the unassayed locus and an assayed locus with known functions [29]. In this study, we set a minimal r^2 threshold of 0.5 in the tagSNP selection, which is an adequately stringent r^2 threshold as recommended by Carlson *et al.* [20]. The haplotype inference, however, was not based on biological plausibility of the SNPs and the haplotype-based analysis came from the multiple comparisons, and the validity of information derived from such haplotypes warrants further studies.

In the single locus analyses, we found that the protective effect appears to be significant in the heterozygotes but not in homozygotes. It is a common phenomenon that the heterozygotes appear to have the greatest risk, compared with other genotype groups, in molecular epidemiological association studies with genotyping data in the literature. Although current knowledge does not provide a convincing explanation for such a finding, there are several possibilities that may lead to the seemingly peculiar finding. In a strictly genetic sense without any selection bias, such data apparently well fit a codominant model, in which two different alleles in combination would have a greater effect than any one of the two that alone would have a weak effect. Molecularly, this is consistent with a hypothesis that if the protein performs its function in the form of a protein dimer with a required symmetric structure, the expression from two different alleles may compromise such a molecule structure requirement, which is often associated with either a reduced or enhanced function of the protein of the dominant allele. This hypothesis needs to be further tested in in-depth molecular mechanistic studies in the future. It is also possible that the variant allele may have a very strong dominant effect so that there is little difference between the effects of the variant homozygotes and heterozygotes.

Statistically, because of the relatively small numbers of the variant homozygotes observed in both cases and controls, the effect of the variant homozygotes might more likely be subject to any selection bias or other unfavorable genotypes than that of the heterozygotes with much larger observations, or simply there is not enough statistical power to detect any real effect among the variant homozygotes. This possibility can only be corrected by much larger studies in the future. As cancer is a disease involving multiple SNPs in multiple genes, a haplotype-based approach might be more powerful than the approach of analyzing a single allele or locus. Statistically, however, such a haplotype-based approach requires a much larger sample size. Therefore, our study of more than 1000 cases and 1000 controls still does not have enough statistical power to detect a real haplotype effect and thus cannot exclude the possibility that some of the results may be due to chance. Technically, however, such a bias towards the heterozygotes could result from a systematic error, particularly in the genotyping using restriction enzymes, but this is unlikely to occur in the TaqMan assay for genotyping using variant-specific probes, as we used in this study.

Several inherited limitations in our study need to be addressed. First of all, the response rate of the cases and controls were 77.8 and 81.3%, respectively, and only 86.2% of the participants were successful for genotyping at the rs1007616 locus, which may have caused some bias possibly because of variation in the sequence targeted by the primers. The general demographic characteristics and smoking habits of the individuals who were and were not included in the final analysis, however, were similar, and the rs1007616 locus was also in high LD with the variant rs3212948 ($D' = 0.882$, $r^2 = 0.756$), which may have captured information that may be lost in the untyped samples; therefore, there would not be substantial bias, if any. Second, the sample size of our study may neither be large enough to detect a small effect from very low penetrance SNPs nor to evaluate gene–environment interactions adequately. Third, our study lacked the information on occupational and environmental exposure in addition to smoking. By matching the controls to the cases on age, sex, and residential area, this kind of bias could, however, be minimized. Finally, the exact biological and functional consequence of these *ERCC1* variants was not known. Therefore, further studies are needed to determine how these tagSNPs may influence the protein function and thus lead to carcinogenesis.

In conclusion, in this case–control study in a Chinese population, we provided evidence that two common variants, rs3212948 (G/C) and rs1007616 (C/T), of *ERCC1* may contribute to the susceptibility of lung cancer. Studies with ethnically diverse populations and functional evaluation are warranted to confirm our findings and to further elucidate the significance of these variants in the development of lung cancer.

Acknowledgements

This work was supported in part by the China National Key Basic Research Program Grants 2002CB512902 (to D. Lu and H. Shen), 2002BA711A10 and 2004CB518605 (to W. Huang), National Outstanding Youth Science Foundation of China 30425001 (to H. Shen), and National '211' Environmental Genomics Grant (to D. Lu). The authors would like to thank Dr Yijiang Chen (The First affiliated Hospital of Nanjing Medical University), Dr Lin Xu (Jiangsu Cancer Hospital) for their assistance in recruiting the patients, Yanhong Liu (Fudan University), Juying Niu (Nanjing Medical University), and personnel from Tongji Medical College for their help in sample management, and the associates from

Department of Genetics, Chinese National Human Genome Center at Shanghai for their technical support.

References

1 Tyczynski JE, Bray F, Parkin DM. Lung cancer in Europe in 2000: epidemiology, prevention, and early detection. *Lancet Oncol* 2003; **4**:45–55.

2 Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol Biomarkers Prev* 2002; **11**:1513–1530.

3 Vispe S, Yung TM, Ritchot J, Serizawa H, Satoh MS. A cellular defense pathway regulating transcription through poly (ADP-ribosylation) in response to DNA damage. *Proc Natl Acad Sci USA* 2000; **97**:9886–9891.

4 Friedberg EC. How nucleotide excision repair protects against cancer. *Nat Rev Cancer* 2001; **1**:22–33.

5 Wei Q, Cheng L, Amos CI, Wang LE, Guo Z, Hong WK, *et al.* Repair of tobacco carcinogen-induced DNA adducts and lung cancer risk: a molecular epidemiologic study. *J Natl Cancer Inst* 2000; **92**:1764–1772.

6 Spitz MR, Wei Q, Dong Q, Amos CI, Wu X. Genetic susceptibility to lung cancer: the role of DNA damage and repair. *Cancer Epidemiol Biomarkers Prev* 2003; **12**:689–698.

7 Wilson MD, Ruttan CC, Koop BF, Glickman BW. ERCC1: a comparative genomic perspective. *Environ Mol Mutagen* 2001; **38**:209–215.

8 Hsia KT, Millar MR, King S, Selfridge J, Redhead NJ, Melton DW, *et al.* DNA repair gene Ercc1 is essential for normal spermatogenesis and oogenesis and for functional integrity of germ cell DNA in the mouse. *Development* 2003; **130**:369–378.

9 Westerveld A, Hoeijmakers JH, van Duin M, de Wit J, Odijk H, Pastink A, *et al.* Molecular cloning of a human DNA repair gene. *Nature* 1984; **310**:425-429.

10 Melton DW, Ketchen AM, Nunez F, Bonatti-Abbondandolo S, Abbondandolo A, Squires S, *et al.* Cells from ERCC1-deficient mice show increased genome instability and a reduced frequency of S-phase-dependent illegitimate chromosome exchange but a normal frequency of homologous recombination. *J Cell Sci* 1998; **111**:395–404.

11 Reed E, Yu JJ, Davies A, Gannon J, Armentrout SL. Clear cell tumors have higher mRNA levels of ERCC1 and XPB than other histological types of epithelial ovarian cancer. *Clin Cancer Res* 2003; **9**:5299–5305.

12 Cheng L, Spitz MR, Hong WK, Wei Q. Reduced expression levels of nucleotide excision repair genes in lung cancer: a case–control analysis. *Carcinogenesis* 2000; **21**:1527–1530.

13 Mohrenweiser HW, Xi T, Vazquez-Matias J, Jones IM. Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. *Cancer Epidemiol Biomarkers Prev* 2002; **11**:1054–1064.

14 Qiao Y, Spitz MR, Guo Z, Hadeyati M, Grossman L, Kraemer KH, *et al.* 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* 2002; **509**:165–174.

15 Matullo G, Peluso M, Polidoro S, Guarrera S, Munnia A, Krogh V, *et al.* Combination of DNA repair gene single nucleotide polymorphisms and increased levels of DNA adducts in a population-based study. *Cancer Epidemiol Biomarkers Prev* 2003; **12**:674–677.

16 Chen P, Wiencke J, Aldape K, Kesler-Diaz A, Miike R, Kelsey K, *et al.* Association of an ERCC1 polymorphism with adult-onset glioma. *Cancer Epidemiol Biomarkers Prev* 2000; **9**:843–847.

17 Sturgis EM, Dahlstrom KR, Spitz MR, Wei Q. DNA repair gene ERCC1 and ERCC2/XPD polymorphisms and risk of squamous cell carcinoma of the head and neck. *Arch Otolaryngol Head Neck Surg* 2002; **128**:1084–1088.

18 Zhou W, Liu G, Park S, Wang Z, Wain JC, Lynch TJ, *et al.* Gene-smoking interaction associations for the ERCC1 polymorphisms in the risk of lung cancer. *Cancer Epidemiol Biomarkers Prev* 2005; **14**:491–496.

19 Shen M, Berndt SI, Rothman N, Demarini DM, Mumford JL, He X, *et al.* Polymorphisms in the DNA nucleotide excision repair genes and lung cancer risk in Xuan Wei, China. *Int J Cancer* 2005; **116**:768–773.

20 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**:106–120.

21 Hu Z, Shao M, Yuan J, Xu L, Wang F, Wang Y, *et al.* Polymorphisms in DNA damage binding protein 2 (DDB2) and susceptibility of primary lung cancer in Chinese: a case–control study. *Carcinogenesis* 2006 [Epub ahead of print].

22 McWhir J, Selfridge J, Harrison DJ, Squires S, Melton DW. Mice with DNA repair gene (ERCC-1) deficiency have elevated levels of p53, liver nuclear abnormalities and die before weaning. *Nat Genet* 1993; **5**:217–224.

23 Dabholkar M, Bostick-Bruton F, Weber C, Bohr VA, Egwuagu C, Reed E. ERCC1 and ERCC2 expression in malignant tissues from ovarian cancer patients. *J Natl Cancer Inst* 1992; **84**:1512–1517.

24 Dabholkar M, Vionnet J, Bostick-Bruton F, Yu JJ, Reed E. Messenger RNA levels of XPAC and ERCC1 in ovarian cancer tissue correlate with response to platinum-based chemotherapy. *J Clin Invest* 1994; **94**:703–708.

25 Metzger R, Leichman CG, Danenberg KD, Danenberg PV, Lenz HJ, Hayashi K, *et al.* ERCC1 mRNA levels complement thymidylate synthase mRNA levels in predicting response and survival for gastric cancer patients receiving combination cisplatin and fluorouracil chemotherapy. *J Clin Oncol* 1998; **16**:309–316.

26 Shirota Y, Stoehlmacher J, Brabender J, Xiong YP, Uetake H, Danenberg KD, *et al.* ERCC1 and thymidylate synthase mRNA levels predict survival for colorectal cancer patients receiving combination oxaliplatin and fluorouracil chemotherapy. *J Clin Oncol* 2001; **19**:4298–4304.

27 Khoury M, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology. Monographs in epidemiology and biostatistics*. New York, New York: Oxford University Press; 1993. pp. 144–145.

28 Kaplan N, Morris R. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol* 2001; **20**:432–457.

29 Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**:1–14.