

Polymorphisms in the Two Helicases *ERCC2/XPD* and *ERCC3/XPB* of the Transcription Factor IIH Complex and Risk of Lung Cancer: A Case-Control Analysis in a Chinese Population

Zhibin Hu,¹ Liang Xu,² Minhua Shao,³ Jing Yuan,⁴ Yi Wang,³ Feng Wang,⁴
Wentao Yuan,² Ji Qian,³ Hongxia Ma,¹ Ying Wang,² Hongliang Liu,³
Weihong Chen,⁴ Lin Yang,² Guangfu Jing,¹ Xiang Huo,¹ Feng Chen,¹
Li Jin,³ Qingyi Wei,⁵ Tangchun Wu,⁴ Daru Lu,³ Wei Huang,² and Hongbing Shen¹

¹Department of Epidemiology and Biostatistics, Cancer Research Center of Nanjing Medical University, Nanjing, China; ²Department of Genetics, Chinese National Human Genome Center at Shanghai; ³State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China; ⁴Institute of Occupational Medicine, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China; and ⁵Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas

Abstract

The transcription factor IIH (TFIIH) helicases *ERCC2/XPD* and *ERCC3/XPB* are responsible for opening the DNA strand around the lesion site during nucleotide excision repair process. Genetic variants in these two genes may be markers for interindividual variability in DNA repair capacity and thus predisposition to cancer risk. In this case-control study of 1,010 incident lung cancer cases and 1,011 age and sex frequency-matched cancer-free controls in a Chinese population, we genotyped eight tagging polymorphisms of *ERCC2* and *ERCC3* using the high-throughput Taqman platform to determine their associations with risk of lung cancer. Although none of the eight polymorphisms was individually associated with lung cancer risk, we found that genetic variants in *ERCC2* and *ERCC3* jointly contributed to lung cancer risk in a dose-response manner. Compared with those with 0 to 1 "at-risk" locus,

subjects carrying >1 at-risk loci were at increased risk for lung cancer [adjusted odds ratio (OR), 1.29; 95% confidence interval (95% CI), 0.98-1.70 for 2 at-risk loci; adjusted OR, 1.38; 95% CI, 1.02-1.85 for 3 at-risk loci; and adjusted OR, 1.51; 95% CI, 1.09-2.10 for ≥ 4 at-risk loci, respectively; $P_{\text{trend}} = 0.015$]. This combined effect was slightly more evident in young subjects (<60 years), males, current smokers, and those with family history of cancer, particularly for histologic type of adenocarcinomas. No evidence for interaction was found. These findings indicate that these tagSNPs of the *ERCC2* and *ERCC3* along with their surrounding regions may serve as biomarkers of susceptibility to lung cancer, which warrant further validation by other population-based and phenotypic studies to determine the biological relevance of these tagSNPs. (Cancer Epidemiol Biomarkers Prev 2006;15(7):1336-40)

Introduction

Tobacco carcinogens induced DNA damage, an important early event in lung carcinogenesis, that is mainly repaired by the nucleotide excision repair (NER) pathway. Mammalian NER has been reconstituted *in vitro*, involving >30 different proteins (1). Among them, helicases *ERCC2/XPD* and *ERCC3/XPB*, two transcription factor IIH (TFIIH) subunits, are responsible for opening the DNA around the site of the lesion, a crucial step to start the NER process (2). Germ line mutations in *ERCC2* and *ERCC3* that alter their protein functions cause severe NER syndromes, such as xeroderma pigmentosum, which represents an extreme low end of the repair spectrum that is associated with a >1,000-fold increased skin cancer risk (3, 4). An ~5-fold

variation in DNA repair capacity (DRC) has been observed in the general population, with reduced DRC shown to be associated with an increased risk of lung cancer (5, 6). It was also suggested that ³¹²Asn and ⁷⁵¹Gln variant genotypes are associated with risk of smoking-related lung cancer (7).

Several *ERCC2* polymorphisms in the coding regions with a relatively high minor allele frequency (MAF) have been identified (8), including two nonsynonymous single nucleotide polymorphisms (nsSNP): G23592A (Asp³¹²Asn) in exon 10 and A35931C (Lys⁷⁵¹Gln) in exon 23. The *ERCC2* ³¹²Asn and ⁷⁵¹Gln variant genotypes were reported to be consistently associated with lower DRC phenotype in genotype-phenotype correlation studies (7, 9). However, it was also found that the ⁷⁵¹Gln genotypes were associated with a higher DRC (10), and that the ³¹²Asn variant genotypes had no effect on DRC (11). Furthermore, a number of case-control studies with different ethnic populations have investigated the associations between *ERCC2* polymorphisms and risk of cancers, particularly smoking-related lung cancer (12). However, the results from these molecular epidemiologic studies are elusive rather than conclusive. In two recently published meta-analyses, elevated lung cancer risks associated with the variant alleles of *ERCC2* ³¹²Asn and ⁷⁵¹Gln were confined to the fixed combination model but not in the random effect model (13, 14), suggesting that these results need to be further evaluated in larger studies within a homogenous population. In addition, these two nsSNPs are not sufficient because they may not be in sufficient linkage disequilibrium (LD) with other untyped causal SNPs in *ERCC2*.

Received 3/20/06; revised 5/3/06; accepted 5/12/06.

Grant support: China National Key Basic Research Program grants 2002CB512902 (D. Lu and H. Shen), 2002CB512905 (T. Wu), 2002BA711A10 (W. Huang), and 2004CB518605 (W. Huang); National Outstanding Youth Science Foundation of China grant 30425001 (H. Shen), National "211" Environmental Genomics grant (D. Lu).

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.

Note: Z. Hu and L. Xu contribute equally to this work.

Requests for reprints: Hongbing Shen, Department of Epidemiology and Biostatistics, Cancer Research Center of Nanjing Medical University, Nanjing 210029, China. Phone: 86-25-8686-2756; Fax: 86-25-8652-7613. E-mail: hbshen@njmu.edu.cn

Co-correspondence to: Wei Huang, Department of Genetics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China. E-mail: huangwei@chgc.sh.cn

Copyright © 2006 American Association for Cancer Research.

doi:10.1158/1055-9965.EPI-06-0194

Table 1. Primary information of selected SNPs/DIPs of ERCC2 and ERCC3 genes

Gene (accession no.) and locus	NCBI rs no.	Position	AA change	Base change	MAF of cases	MAF of controls	% Genotyped
ERCC2/XPD (AY092780) 19q13.3	rs1618536	003642		G/A	0.434	0.450	97.0
	rs1799786	007209		C/T	0.065	0.056	96.5
	rs1799793	007988	D312N	G/A	0.064	0.057	96.8
	rs3916823	010053		AAAA-/-	0.131	0.125	97.6
	rs13181	020337	K751Q	A/C	0.079	0.069	97.6
ERCC3/XPB (AY163769) 2q21	rs2271026	001870		T/C	0.121	0.113	99.1
	rs4150416	005806		T/G	0.004	0.007	96.6
	rs4150441	011440		A/G	0.405	0.426	95.3

NOTE: U.S. Environmental Genome Project SNP database (see <http://egp.gs.washington.edu/directory.html>). Abbreviations: AA, amino acid; MAF, minor allele frequencies.

Because no common (i.e., MAF > 0.05) nsSNPs in the coding region nor functional regulatory variants were identified in ERCC3 to date, there are no studies that have investigated the functional relevance of the SNPs and their associations with cancer risk. Considering similar functions of ERCC2 and ERCC3 proteins in the NER pathway, we determined genotypes of three common tagSNPs of ERCC2 in addition to the two nsSNPs (Asp³¹²Asn and Lys⁷⁵¹Gln) and three tagSNPs in ERCC3 by using high-throughput Taqman genotyping assay in a case-control study of 1,010 incident lung cancer cases and 1,011 age and sex frequency-matched cancer-free controls in a Chinese population. We tested the hypothesis that genetic variants in the ERCC2 and ERCC3 genes are associated with risk of lung cancer.

Materials and Methods

Study Populations. The study population and subject characteristics were previously described elsewhere (15). In brief, a total of 1,299 cases with histopathologically confirmed incident lung cancer were recruited between July 2002 and November 2004 from four hospitals of three metropolitan cities along the Yangzi river, including 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. Of 1,299 subjects approached for recruitment, 1,010 (77.8%) patients consented to participate in the study and provided blood samples (487 from Nanjing, 156 from Shanghai, and 367 from Wuhan). The control subjects consisted of patients with diseases other than cancer recruited from other clinics of the same hospital (425 from Nanjing, 137 from Shanghai, and 449 from Wuhan) during the same time period when the cases were recruited. All the control subjects were frequency matched to the cases on age (± 5 years), sex, and residential area (urban or countryside). Each participant was scheduled for an interview after a written informed consent was obtained, 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 <1 cigarette per day and <1 year in their lifetime were defined as never smokers; otherwise, they were considered as ever smokers. Those ever smokers who quit for >1 year were considered former smokers. Pack-years [(cigarettes per day / 20) \times years smoked] were calculated to indicate the cumulative smoking dose, and the ever smokers were further dichotomized by the cumulative dose of 29 pack-years according to the pack-years distribution of the controls. Family history of cancer was defined as any self-reported cancer in first-degree relatives (parents, siblings, or children). After interview, a one-time 5-mL venous blood sample was collected from each participant. The study was approved by the institutional review boards of Nanjing Medical University, Fudan Univer-

sity, and Tongji Medical College of Huazhong University of Science and Technology, China.

Polymorphism Selection. We used the resequencing data of 90 individuals in the Environmental Genome Project database (<http://egp.gs.washington.edu/>) to select tagSNPs from the reported SNPs/deletion/insertion polymorphisms (DIP) based on their putative functional potentials (i.e., nsSNPs) and from the calculation of pairwise linkage disequilibrium. A greedy algorithm was used to choose the tagSNPs given a minimal LD variable r^2 threshold of 0.5 according to the polymorphism density in ERCC2 and ERCC3 (16). Therefore, if we have a tagSNP list T , the number of SNPs/DIPs identified at level t is counted as $M_T = \text{Count}_t(\text{Max}_{i,j \in T} \text{LD}_{ij} > t)$. By using this selection method, we selected seven tagSNPs, including the functional Asp³¹²Asn and Lys⁷⁵¹Gln variants, for ERCC2 from a total of 39 common SNPs/DIPs (MAF > 0.1), and four tagSNPs for ERCC3 from a total of 36 common SNPs/DIPs available in the Environmental Genome Project database (<http://egp.gs.washington.edu/>). However, because the primers (probes) for some loci could not be successfully constructed by the ABI Primer Express Oligo Design software, we only successfully genotyped five of the seven ERCC2 SNPs/DIPs and three of the four ERCC3 SNPs/DIPs in our 1,010 lung cancer patients and 1,011 cancer-free controls (Table 1).

Laboratory Assays. Genotyping was done by the 5'-nuclease (Taqman) assay, using the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA), in 384-well format, at the 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) and available upon request. The intensity of each SNP should meet the criteria of three clear clusters in two scales generated by SDS software (ABI).

Statistical Analyses. Differences in select demographic variables (smoking status, pack-years smoked, and the frequency of ERCC2 and ERCC3 genotypes between the cases and the controls) were evaluated by using the χ^2 test. The associations between ERCC2 and ERCC3 genotypes and lung cancer risk were estimated by computing the odds ratios (OR) and 95% confidence intervals (95% CI) from both univariate and multivariate logistic regression analyses with adjustment for age, sex, family history of cancer, and pack-years of smoking. To evaluate the effects of the combined genotypes, we first combined the heterozygote and variant homozygote genotype into one genotype for each of seven loci (0 = reference genotype; 1 = risk genotypes) under the assumption of a dominant model for the variant alleles. Then, we combined all dichotomized variables of the seven loci as a new combined genotype variable and categorized this new variable into 0 to 1, 2, 3, and ≥ 4 strata according to the number of "at-risk" loci (i.e., the ones that were more frequent in the cases than in the controls). This categorical variable was coded as dummy variables whose associations

Table 2. ERCC2 and ERCC3 polymorphisms and lung cancer risk in a Chinese population

Genotype	Cases, n (%)	Controls, n (%)	Adjusted OR (95% CI)*
ERCC2 rs1618536	975	985	
AA	177 (18.2)	199 (20.2)	1.00
AG	492 (50.5)	488 (49.5)	1.10 (0.86-1.40)
GG	306 (31.4)	298 (30.3)	1.15 (0.89-1.51)
AG/AA	798 (81.9)	786 (79.8)	1.12 (0.89-1.41)
ERCC2 rs1799786	965	986	
CC	844 (87.5)	877 (89.0)	1.00
CT	117 (12.1)	108 (11.0)	1.12 (0.84-1.49)
TT	4 (0.4)	1 (0.1)	6.18 (0.68-56.52)
CT/TT	121 (12.5)	109 (11.1)	1.16 (0.87-1.54)
ERCC2 rs1799793	970	986	
GG	850 (87.6)	874 (88.6)	1.00
AG	116 (12.0)	111 (11.3)	1.06 (0.80-1.41)
AA	4 (0.4)	1 (0.1)	6.13 (0.67-56.11)
AG/AA	120 (12.4)	112 (11.4)	1.10 (0.83-1.46)
ERCC2 rs3916823	986	987	
AAAA/AAAA	744 (75.5)	750 (76.0)	1.00
AAAA/-	225 (22.8)	227 (23.0)	1.01 (0.81-1.25)
-/-	17 (1.7)	10 (1.0)	1.88 (0.84-4.21)
AAAA/- + -/-	242 (24.5)	237 (24.0)	1.04 (0.84-1.29)
ERCC2 rs13181	975	997	
AA	827 (84.8)	865 (86.8)	1.00
AC	141 (14.5)	127 (12.7)	1.16 (0.89-1.51)
CC	7 (0.7)	5 (0.5)	1.64 (0.50-5.36)
AC/CC	148 (15.2)	132 (13.2)	1.18 (0.91-1.53)
ERCC3 rs2271026	1,002	1,000	
TT	774 (77.3)	787 (78.7)	1.00
CT	214 (21.4)	200 (20.0)	1.07 (0.86-1.34)
CC	14 (1.4)	13 (1.3)	1.02 (0.46-2.24)
CT/CC	228 (22.8)	213 (21.3)	1.07 (0.86-1.33)
ERCC3 rs4150441	952	975	
AA	149 (15.7)	178 (18.2)	1.00
AG	474 (49.8)	475 (48.8)	1.20 (0.93-1.56)
GG	329 (34.6)	322 (33.1)	1.28 (0.97-1.69)
AG/GG	803 (84.4)	797 (81.9)	1.24 (0.97-1.58)
Combined genotypes	847	899	
0-1 at-risk locus	136 (16.1)	186 (20.7)	1.00
2 at-risk loci	331 (39.1)	349 (38.8)	1.29 (0.98-1.70)
3 at-risk loci	231 (27.3)	230 (25.6)	1.38 (1.02-1.85)
≥4 at-risk loci	149 (17.6)	134 (14.9)	1.51 (1.09-2.10)
			$P_{\text{trend}} = 0.02$

*Adjusted for age, sex, pack-years of smoking, and family history of cancer.

with cancer risk were evaluated by them into a multivariate logistic regression model with and without adjustment for other covariates. The association between combined ERCC2 and ERCC3 variant genotypes (≥ 2 at-risk loci versus 0-1 at-risk locus) and lung cancer risk was also evaluated by stratification analyses by variables of interest, such as age, sex, smoking status, family history of cancer, and histologic types. The potential gene-environment interaction was evaluated in SAS PROC LOGIST by using a likelihood ratio test. In addition, we used the PHASE 2.0 Bayesian algorithm (17) to infer haplotype/diplotype frequencies based on the observed genotypes for each gene. Diplotype was the most probable haplotype pair for each individual. Unconditional logistic regression analyses were conducted to estimate ORs and 95% CIs for participants carrying either 1 or 2 versus 0 copies of each common haplotype (MAF ≥ 0.05) for the trichotomized diplotypes. All the statistical analyses were done with Statistical Analysis System software (v.8.0e; SAS Institute, Cary, NC).

Results

The distributions of selected characteristics between lung cancer patients and controls were previously described elsewhere (15). Overall, our frequency matching on age and sex was adequate ($P = 0.98$ for age and $P = 0.30$ for sex).

Smoking, self-reported family history of cancer in the first-degree relatives were significant risk factors for lung cancer. Of the 1,010 cancer patients, 430 (42.6%) were adenocarcinoma, 335 (33.2%) were squamous cell carcinoma, 65 (6.4%) were small cell carcinoma, and 180 (17.8%) were large cell, mixed cell, or undifferentiated carcinomas.

All genotype distributions in the controls were consistent with those expected from the Hardy-Weinberg equilibrium. However, one SNP (i.e., rs4150416 in ERCC3) had the MAF < 0.01 in both cases and controls in this Chinese population and therefore were excluded from subsequent analyses. About half of the SNPs/DIPs in this study population represent a MAF 10% lower than those reported in the Environmental Genome Project SNP database (<http://egp.gs.washington.edu>), which may reflect either ethnic differences or frequency bias due to small sample sizes from which the database derived.

As shown in Table 2, none of the SNPs/DIPs in both ERCC2 and ERCC3 was associated with lung cancer risk in the single-locus analysis. Therefore, we did combined analyses, assuming a dominant model (i.e., to combine the heterozygous and homozygous at-risk genotypes versus the genotype with no at-risk allele) for each locus. A total of 847 (83.9%) cases and 899 (88.9%) controls had been successfully genotyped for all the seven loci. Compared with those who had 0 to 1 at-risk locus, subjects carrying > 1 at-risk loci had increased risks of lung cancer (adjusted OR, 1.29; 95% CI, 0.98-1.70 for 2 at-risk loci; adjusted OR, 1.38; 95% CI, 1.02-1.85 for 3 at-risk loci; and adjusted OR, 1.51; 95% CI, 1.09-2.10 for ≥ 4 at-risk loci). A significant locus dose-response effect on lung cancer risk was also observed ($P_{\text{trend}} = 0.02$). In stepwise logistic regression analysis, when we set the threshold P for one variant to entry the model as 0.20, rs4150441 in ERCC3 and rs13181 in ERCC2 were the two polymorphisms included in the model. The ORs (95% CI) were 1.31 (1.01-1.71) for rs4150441 and 1.20 (0.91-1.58) for rs13181 (data not shown).

The association between combined ERCC2 and ERCC3 variant genotypes (i.e., ≥ 2 at-risk loci versus 0-1 at-risk locus) and lung cancer risk was further evaluated by stratification analyses by age, sex, smoking status, family history of cancer, and histologic types (Table 3). As shown in Table 3, the effects of combined variant genotypes were slightly more evident in young subjects (≤ 60 years; adjusted OR, 1.58; 95% CI, 1.10-2.28), males (adjusted OR, 1.45; 95% CI, 1.09-1.95), current smokers (adjusted OR, 1.41; 95% CI, 0.96-2.80), subjects with family history of cancer (adjusted OR, 1.91; 95% CI, 0.98-3.74), and subjects with histologic type of adenocarcinomas (adjusted OR, 1.61; 95% CI, 1.14-2.26). However, we found no evidence for any interaction between combined variant genotypes and smoking/family cancer history in the multivariate logistic regression model (data not shown).

In addition, we did haplotype/diplotype inference using the PHASE 2.0 program based on the known genotypes of the two NER genes. For ERCC2, no obvious association was obtained for both haplotype and diplotype analyses (data not shown). For ERCC3, compared with the haplotype TA carrying no at-risk allele, the TG haplotype was associated with a 1.12-fold increased risk of lung cancer (95% CI, 0.98-1.29), and the CA haplotype was associated with a 1.18-fold elevated risk (95% CI, 0.95-1.46). Table 4 summarizes the associations between ERCC3 common diplotypes and lung cancer risk. Compared with the diplotype carrying 0 copy of TG haplotype, the dipolotype carrying 1 copy of TG haplotype and 2 copies of the TG haplotypes were associated, respectively, with a 1.25-fold (95% CI, 0.97-1.61) and a 1.28-fold (95% CI, 0.97-1.68) increased risks of lung cancer. Similarly, the dipolotype carrying 1 copy of TA haplotype and 2 copies of the TA haplotypes were associated, respectively, with a 13% (95% CI, 0.72-1.06) and a 24% (95% CI, 0.55-1.04) decreased lung cancer risks (Table 4).

Table 3. Stratified analyses between the combined *ERCC2* and *ERCC3* genotypes and lung cancer risk in a Chinese population

	<i>ERCC2</i> and <i>ERCC3</i> combined genotypes					
	Cases (<i>n</i> = 847), <i>n</i> (%)		Controls (<i>n</i> = 899), <i>n</i> (%)		Adjusted OR (95% CI)*	
	0-1 at-risk locus	≥2 at-risk loci	0-1 at-risk locus	≥2 at-risk loci	0-1 at-risk locus	≥2 at-risk loci
Total	136 (16.1)	711 (83.9)	186 (20.7)	713 (79.3)	1.00	1.36 (1.06-1.75)
Age (y)						
≤60	60 (14.9)	344 (85.2)	91 (21.0)	342 (79.0)	1.00	1.58 (1.10-2.28)
>60	76 (17.2)	367 (82.8)	95 (20.4)	371 (79.6)	1.00	1.19 (0.84-1.68)
Sex						
Male	103 (15.5)	560 (84.5)	138 (20.9)	522 (79.1)	1.00	1.45 (1.09-1.95)
Female	33 (17.9)	151 (82.1)	48 (20.1)	191 (79.9)	1.00	1.10 (0.67-1.80)
Smoking status						
Current smokers	59 (17.5)	278 (82.5)	73 (22.7)	249 (77.3)	1.00	1.41 (0.96-2.08)
Former smokers	37 (14.6)	217 (85.4)	27 (18.6)	118 (81.4)	1.00	1.37 (0.79-2.38)
Nonsmokers	40 (15.6)	216 (84.4)	86 (19.9)	346 (80.1)	1.00	1.24 (0.82-1.89)
Family history of cancer						
No	115 (16.5)	582 (83.5)	159 (20.3)	623 (79.7)	1.00	1.28 (0.98-1.68)
Yes	21 (14.0)	129 (86.0)	27 (23.1)	90 (76.9)	1.00	1.91 (0.98-3.74)
Histologic types						
Adenocarcinomas	51 (13.8)	319 (86.2)			1.00	1.61 (1.14-2.26)
Squamous cell	51 (17.8)	236 (82.2)			1.00	1.22 (0.85-1.75)
Small cell	9 (16.4)	46 (83.6)			1.00	1.40 (0.67-2.94)
Other carcinomas [†]	25 (18.5)	110 (81.5)	186 (20.7)	713 (79.3)	1.00	1.10 (0.41-1.97)

*Adjusted for age, sex, family history of cancer, and pack-years of smoking.

[†]Other carcinomas means the large cell, mixed cell carcinomas or undifferentiated carcinoma.

Discussion

In this lung cancer case-control study in a Chinese population, we found, for the first time, that genetic variants in *ERCC2* and *ERCC3*, each conferring a small portion of risk, may jointly contribute to lung cancer risk. These findings indicate that some representative tagSNPs of *ERCC2* and *ERCC3* along with their surrounding regions may serve as biomarkers of susceptibility to lung cancer. However, these findings warrant further validation by studies on biological significance of these SNPs/DIPs and population-based association studies.

In the current study, we used a polymorphism selection strategy combining both hypotheses-driven (nsSNPs) and hypotheses-free (tagSNPs) approaches to identify the most representative SNPs in *ERCC2* and *ERCC3*. Although we had a few missing data in each locus genotyped, the relatively large sample size made it possible to perform both single-locus evaluation and combined analyses. Furthermore, because the 90 U.S. individuals in the Environmental Genome Project database had mixed ethnic background, and because Asian populations (including Chinese) were defined as that with lower haplotype diversity and higher pairwise LD compared with other populations (18), we thought our selected tagSNPs from the Environmental Genome Project SNP database would be more informative than we originally estimated. For example, for the *ERCC2* gene, the pairwise LD between rs1799786 and rs1799793 in our study population was 0.90, which was higher than 0.48 in the NIEHS SNP database, suggesting that genetic variants of these genes in Asian population might have the more power to capture variations across the gene. Although we only geno-

typed two informative SNPs in *ERCC3* (rs2271026 and rs4150441), these two SNPs were exactly the two of the three tagSNPs that the HapMap project recommended for Chinese population at a minimal LD variable r^2 threshold 0.8 (<http://www.hapmap.org/>). Furthermore, all the captured SNPs by the other tagSNPs in HapMap database had a pairwise r^2 of 0.657 with the typed rs4150441 SNP; therefore, the two typed SNPs in our study are the actual tagSNPs for *ERCC3* in the Chinese population according to the HapMap database, although this information was not available when the project started.

The exact biological mechanisms of how these *ERCC2* and *ERCC3* SNPs/DIPs affect cancer risk at the molecular level remain to be unraveled. However, published studies on the structure and functions of the two nsSNPs in *ERCC2* are informative in understanding the potential roles of these polymorphisms. The *ERCC2* Lys⁷⁵¹Gln polymorphism is about 50 bases upstream from the poly(A) signal and therefore may alter XPD protein function. Apart from the direct link between genotype and DRC phenotype (7, 9), the *ERCC2* Asp³¹²Asn and Lys⁷⁵¹Gln variants have been suggested to be associated with higher levels of chromosomal aberrations induced by X-ray (11) and increased levels of DNA adducts (19, 20). However, compared with these "intermediates," cancer as an end point might be more complicated in the presence of many other unknown competing risk factors, and a single genetic variant is insufficient to predict the complex phenotype under the polygenic model, where a large number of alleles, each conferring a small genotypic effect (perhaps of the order of 1.1-1.5), may combine additively or multiplicatively to confer a range of cancer risk in the general population (21, 22).

Table 4. *ERCC3* common diplotypes and lung cancer risk

Diploypes*	0 copy		1 copy		2 copies	
	Case/control	OR	Case/control	OR (95% CI) [†]	Case/control	OR (95% CI) [†]
<i>ERCC3</i> / <i>XPB</i>						
TG	151/179	1.00	530/512	1.25 (0.97-1.61)	329/320	1.28 (0.97-1.68)
TA	507/470	1.00	416/439	0.87 (0.72-1.06)	87/102	0.76 (0.55-1.04)
CA	782/801	1.00	215/197	1.10 (0.88-1.38)	13/13	0.94 (0.42-2.10)

*Polymorphic bases were in 5' to 3' order as listed in Table 2.

[†]Adjusted for age, sex, family history of cancer, and pack-years of smoking.

In the present study, we observed that the frequency of histologic types of lung cancer is different from what were found in other Asian countries probably because of the different smoking rate of the populations, different proportion of females in lung cancer cases, different proportion of the cases from countryside, different diagnostic criteria of different countries, and potential selection bias of different hospitals. Differences in allele frequencies or LD structure across different populations or ethnic groups may be another concern for possible confounding effects because the same high-risk allele may have a very different pattern of association with marker alleles and haplotypes in different ethnic groups (23, 24). Prevalence of the variant ERCC2 alleles and genotypes varies markedly with ethnicity (7, 20, 25-31). In the current study, allele frequencies of the ERCC2³¹²Asn and ERCC2⁷⁵¹Gln were both consistent with those reported by Liang et al. in a large case-control study of 1,006 lung cancer cases and 1,020 controls in North China (28).

Like all other case-control studies, however, inherited biases in the present study may have led to spurious findings. First of all, because our study was a hospital-based study, our control subjects may not be representative of the general population. However, we believe that our results are unlikely to be attributable to selection bias because we used a relatively large number of incident lung cancer cases and matched the controls to the cases on age, sex, and residential area. Second, although we included >1,000 cases and 1,000 controls, we found that none of the variant genotypes was associated with a significantly increased or decreased risk of lung cancer in the single-locus analysis, which may reflect the need of larger sample sizes for association studies of these low-penetrance genes. Clearly, our sample size was not large enough to identify significant associations of the effect in different strata in subgroup analyses, and we also failed to evaluate gene-environment interactions adequately. However, the approach of using combined tagSNPs may represent an alternative way of analyzing the overall effect of the ERCC2/ERCC3 genetic variants as well as potential joint effect between these two genes. Because we focused on the overall effect of the combined genotypes and the dose-response effect for the association between the number of at-risk loci and lung cancer risk, no specific combined genotype could be specified for the observed effect. Because the combined genotypes were based on the tagSNPs (other than functional SNPs), these should be free of the any assumption on their functionality that is difficult to be investigated in the laboratories. Third, for each locus, the genotyping failure rate was <5%, but the overall genotype failure rate for all seven loci combined was 12% for the controls and 16% for the cases. However, there was no statistically significant difference in the distributions of demographic characteristics and smoking habits of those who were and were not included in the final analysis. Furthermore, because of our limited genotyping efforts due to our financial constraint, the five SNPs in ERCC2 and three in ERCC3 we genotyped were not enough to capture or represent all the genetic variants of these two genes. Finally, except for tobacco smoking, other factors, such as occupational exposure, and certain dietary components, might interact with ERCC2/ERCC3 genotypes or act as potential confounders. Unfortunately, information on these factors in our case-control study was not available. It would be interesting to investigate interactions between ERCC2/ERCC3 genotypes and these risk factors in future studies. In addition, because we only used statistical methods to infer the haplotypes/diplootypes instead of having actual phase information from family studies, the results need further validation.

In conclusion, our current study, benefited from a relatively large, homogenous ethnic Han Chinese population and simultaneous evaluation of multiple tagSNPs covering common variants in the two helicase genes of the TFIIH complex

involved in the NER pathway, provided a snapshot of the relationship between ERCC2 and ERCC3 variants and lung cancer susceptibility. Validation of these findings in larger studies of other populations is needed.

References

- Friedberg EC. How nucleotide excision repair protects against cancer. *Nat Rev Cancer* 2001;1:22-33.
- Egly JM. The 14th Datta Lecture. TFIIH: from transcription to clinic. *FEBS Lett* 2001;498:124-8.
- Taylor EM, Broughton BC, Botta E, et al. Xeroderma pigmentosum and trichothiodystrophy are associated with different mutations in the XPD (ERCC2) repair/transcription gene. *Proc Natl Acad Sci U S A* 1997;94:8658-63.
- Coin F, Marinoni JC, Rodolfo C, Fribourg S, Pedrini AM, Egly JM. Mutations in the XPD helicase gene result in XP and TTD phenotypes, preventing interaction between XPD and the p44 subunit of TFIIH. *Nat Genet* 1998;20:184-8.
- Wei Q, Cheng L, Amos CI, et al. Repair of tobacco carcinogen-induced DNA adducts and lung cancer risk: a molecular epidemiologic study. *J Natl Cancer Inst* 2000;92:1764-72.
- Shen H, Spitz MR, Qiao Y, et al. Smoking, DNA repair capacity and risk of nonsmall cell lung cancer. *Int J Cancer* 2003;107:84-8.
- Spitz MR, Wu X, Wang Y, et al. Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients. *Cancer Res* 2001;61:1354-7.
- Shen MR, Jones IM, Mohrenweiser H. Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans. *Cancer Res* 1998;58:604-8.
- Qiao Y, Spitz MR, Guo Z, 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;509:165-74.
- Moller P, Knudsen LE, Frenzt G, Dybdahl M, Wallin H, Nexø BA. Seasonal variation of DNA damage and repair in patients with non-melanoma skin cancer and referents with and without psoriasis. *Mutat Res* 1998;407:25-34.
- Lunn RM, Helzlsouer KJ, Parshad R, et al. XPD polymorphisms: effects on DNA repair proficiency. *Carcinogenesis* 2000;21:551-5.
- Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol Biomarkers Prev* 2002;11:1513-30.
- Hu Z, Wei Q, Wang X, Shen H. DNA repair gene XPD polymorphism and lung cancer risk: a meta-analysis. *Lung Cancer* 2004;46:1-10.
- Benhamou S, Sarasin A. ERCC2/XPD gene polymorphisms and lung cancer: a HuGE review. *Am J Epidemiol* 2005;161:1-14.
- Hu Z, Shao M, Yuan J, et al. Polymorphisms in DNA damage binding protein 2 (DDB2) and susceptibility of primary lung cancer in Chinese: a case-control study. *Carcinogenesis* 2006 Mar 7; [Epub ahead of print].
- 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;74:106-20.
- Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003;73:1162-9.
- Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002;296:2225-9.
- Matullo G, Palli D, Peluso M, et al. XRCC1, XRCC3, XPD gene polymorphisms, smoking and (32)P-DNA adducts in a sample of healthy subjects. *Carcinogenesis* 2001;22:1437-45.
- Hou SM, Falt S, Angelini S, et al. The XPD variant alleles are associated with increased aromatic DNA adduct level and lung cancer risk. *Carcinogenesis* 2002;23:599-603.
- Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037-48.
- Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000;405:847-56.
- Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001;69:1-14.
- Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 2002;11:505-12.
- Zhou W, Liu G, Miller DP, et al. Gene-environment interaction for the ERCC2 polymorphisms and cumulative cigarette smoking exposure in lung cancer. *Cancer Res* 2002;62:1377-81.
- Butkiewicz D, Rusin M, Enewold L, Shields PG, Chorazy M, Harris CC. Genetic polymorphisms in DNA repair genes and risk of lung cancer. *Carcinogenesis* 2001;22:593-7.
- Misra RR, Ratnasinghe D, Tangrea JA, et al. Polymorphisms in the DNA repair genes XPD, XRCC1, XRCC3, and APE/ref-1, and the risk of lung cancer among male smokers in Finland. *Cancer Lett* 2003;191:171-8.
- Liang G, Xing D, Miao X, et al. Sequence variations in the DNA repair gene XPD and risk of lung cancer in a Chinese population. *Int J Cancer* 2003;105:669-73.
- David-Beabes GL, Lunn RM, London SJ. No association between the XPD (Lys⁷⁵¹Gln) polymorphism or the XRCC3 (Thr²⁴¹Met) polymorphism and lung cancer risk. *Cancer Epidemiol Biomarkers Prev* 2001;10:911-2.
- Park JY, Lee SY, Jeon HS, et al. Lys⁷⁵¹Gln polymorphism in the DNA repair gene XPD and risk of primary lung cancer. *Lung Cancer* 2002;36:15-6.
- Hamajima N, Saito T, Matsuo K, et al. Genotype frequencies of 50 polymorphisms for 241 Japanese non-cancer patients. *J Epidemiol* 2002;12:229-36.