Genetic predisposition to lung adenocarcinoma among never-smoking Chinese with different epidermal growth factor receptor mutation status

Li Han [a], Cheuk-Kwong Lee [b], Herbert Pang [c], Hong-Tou Chan [d], Iek-Long Lo [d], Sze-Kwan LAM [a], Tak-Hong Cheong [d], James Chung-Man Ho [a]

[a] Division of Respiratory Medicine, Department of Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong SAR

[b] Hong Kong Red Cross Blood Transfusion Service, Hong Kong SAR

[c] School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR

[d] Pneumology Department, Centro Hospitalar C.S. Januario Macau, Macau

Corresponding author:

Dr. James Chung-man HO M.D. FRCP,

Department of Medicine, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong SAR, China.

Tel: (852) 2255 4999; Fax: (852) 2872 5828; Email: jhocm@hku.hk

Keywords: Lung adenocarcinoma; epidermal growth factor receptor mutation; single nucleotide polymorphisms; gene-environment interaction; never-smokers

Conflict of interest: No author reports any potential conflicts of interest.

Word count: 3495

Total number of figures: 3

Total number of supplementary figures: 1

Total number of tables: 3

Total number of supplementary tables: 6

Abstract:

Objectives:

The inconsistent findings from genetic association studies may be related to the heterogeneity in different molecular subtypes of lung cancer. This study evaluated the predisposing single-nucleotide polymorphisms (SNPs) in epidermal growth factor receptor (EGFR) mutant and EGFR wild-type lung adenocarcinoma separately among never-smokers.

Materials and Methods:

This was a two-stage case-control study. Never-smokers with pathologically confirmed lung adenocarcinoma and healthy controls were recruited in Hong Kong and Macau. Genomic DNA was extracted and genotyped by MassARRAY. In the discovery stage, 51 SNPs were investigated at the SNP, gene and pathway level among 103 EGFR mutant and 78 EGFR wild-type lung adenocarcinoma cases compared with matched controls. In the validation stage, SNPs that were identified with significant lung cancer risk were replicated in a separate cohort of 84 lung adenocarcinoma cases and compared with 103 Chinese Han, Beijing and 105 Chinese Han, Southern public controls from the 1000 genome database.

Results and Conclusion:

The genetic association of IL-6 rs2069840 with EGFR mutant lung adenocarcinoma was ascertained. In the discovery stage, haplotype GGG in three SNPs (rs2069840, rs2069852, rs2066992) of IL-6, synergetic effects of IL-6 rs2069840 and environmental tobacco smoke in the workplace were found to be related to EGFR mutant lung adenocarcinoma. ERCC2 rs238406 showed a marginally significant association with EGFR mutant lung adenocarcinoma in the validation stage (P=0.096). ERCC2 rs50871 and ATM rs611646 showed significant association with EGFR wild-type lung

adenocarcinoma in the discovery stage. In conclusion, IL-6 rs2069840 conferred susceptibility to EGFR mutant lung adenocarcinoma in a Hong Kong and Macau never-smoking Chinese population.

(Word count: 256 words)

Highlights

- Genetic association of IL-6 rs2069840 with EGFR mutant lung adenocarcinoma
- Synergism of IL-6 rs2069840 and environmental tobacco smoke in the workplace
- Association of ERCC2 rs50871 with EGFR wild-type lung adenocarcinoma
- Association of ATM rs611646 with EGFR wild-type lung adenocarcinoma

Abbreviations:

SNPs: single-nucleotide polymorphisms; EGFR: epidermal growth factor receptor; NSCLC: non-small cell lung cancer; ADC: lung adenocarcinoma; LCINS: lung cancer in never-smokers; ALK: anaplastic lymphoma kinase; GWAS: genome-wide association studies; HKU/HA HKW IRB: Institutional Review Board of the University of Hong Kong/ Hospital Authority Hong Kong West Cluster; CHB: Chinese Han, Beijing; CHS: Chinese Han, Southern; MAF: minor allele frequency; SD: standard deviation; LD: linkage disequilibrium; MDR: multifactor dimensionality reduction; AIC: Akaike information criterion; TA: testing accuracy; CVC: cross-validation consistency; FDR: false discovery rate; HWE: Hardy-Weinberg equilibrium; TKI: tyrosine-kinase inhibitor; NER: nucleotide excision repair; FS: functional significance.

Financial support: This research received no grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Introduction

Lung adenocarcinoma (ADC) has become the predominant cell type in lung cancer cases throughout the world [1]. Although tobacco smoking is a dominant environmental risk factor for development of lung ADC, lung cancer in never-smokers (LCINS) has emerged as a distinct disease entity in terms of biological behaviour, molecular profile, and therapeutic options [2, 3]. The predilection of epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) re-arrangement among LCINS has led to the rapid development of targeted therapies for lung ADC. Nonetheless EGFR mutations are found in almost 80% of never-smoking lung ADC in the Chinese population [4]. Therefore, even among never-smokers, lung ADC is now considered a heterogeneous disease with different molecular characteristics. There has been great interest in the identification of potential genes that are associated with a predisposition to development of LCINS. This may allow early recognition of never-smokers who have an increased susceptibility to NSCLC. Candidate gene studies and genome-wide association studies (GWAS) have identified a link between single nucleotide polymorphisms (SNPs) of genes involved in DNA repair, inflammation, carcinogen metabolism and tumour suppression and the development of lung cancer among never-smokers. The inconsistent results from previous reports may have been due to different genotyping methods [5, 6], population demographics, family history of cancers and environmental exposure [6, 7]. Nonetheless the differences among lung ADC with distinct driver oncogenes are likely crucial. This study was designed with the primary objective to investigate the predisposing SNPs in EGFR mutant and wild-type

lung ADC among never-smokers. The secondary objectives included identification of environmental risk factors as well as possible gene-environment interactions that contribute to the development of EGFR mutant and wild-type lung ADC.

2. Materials and methods

2.1. Subject recruitment

This was a two-stage (discovery and validation stage) case-control study. The study was approved by the Institutional Review Board of the University of Hong Kong/ Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) (UW13-343) and the hospital ethics committee of the Centro Hospitalar C.S. Januario, Macau. Eligible ethnically Chinese participants with confirmed primary lung cancer were prospectively recruited at Queen Mary Hospital in Hong Kong starting from September 2006 to a HKU/HA HKW IRB-approved project to compile a prospective lung cancer database (UW06-151 T/1176). Cases and controls with Chinese ethnicity, age over 18 years, male or female, who had never smoked or had smoked fewer than 100 cigarettes during their lifetime were included. Cases with confirmed primary lung adenocarcinoma, with or without EGFR mutations were included. Those with a history of cancer other than lung in origin, and controls with a history of any cancer or respiratory disease were excluded. By February 2015, a total of 653 lung cancer patients had been recruited from Hong Kong and Macau, regardless of histological type, tumour molecular subtype or smoking status, of whom 299 never-smoking patients with lung adenocarcinoma served as cases: 146 (61%) EGFR mutants, 93 (39%) EGFR wild-type, 60 unknown EGFR status

(54 were diagnosed before 2011 when the EGFR test was unavailable, 6 with insufficient samples for testing). The EGFR mutation test was performed by standard methodology, either by direct sequencing or allele-specific polymerase chain reaction, depending on the quality of tumour sample. 453 blood donors were recruited from the Hong Kong Red Cross from May 2013 to February 2015 of whom 332 never-smoking healthy donors were chosen as controls. Controls were randomly selected and individually matched with cases by gender and age ± 5 years in a 1:1 ratio. After matching, 103 EGFR mutant and 78 EGFR wild-type age- and gender-matched pairs were included in the discovery stage. In the validation stage, a separate cohort of 84 never-smoking Chinese patients with EGFR mutant lung adenocarcinoma recruited from Hong Kong and Centro Hospitalar C.S. Januario, Macau from March 2015 to December 2016 were chosen as independent cases, while 103 Chinese Han, Beijing (CHB) and 105 Chinese Han, Southern (CHS) public controls in 1000 genome database were selected as controls [9].

2.2. SNP selection, sample preparation and genotyping

Fifty-one SNPs in 14 genes belonging to four different pathways (DNA repair, carcinogen metabolism, inflammation, tumour suppression) were genotyped (Supplementary Table 1). The SNPs were selected based on the following criteria: (1) SNPs with known or putative functions from previously reported candidate genetic association studies or GWAS; (2) Tagger SNPs with pairwise linkage disequilibrium (LD) set as a squared correlation coefficient (r^2) more than 0.8 ($r^2 \geq 0.8$); (3) Minor allele

frequency (MAF) ≥ 5% in CHS descendants or CHB descendant-based data from 1000 genome project [9]. A venous blood sample (10mLs) was taken from all subjects, with buffy coat separated by centrifugation and stored at -20^{o}C. Genomic DNA was extracted from the stored buffy coat using a DNA Blood Kit (Qiagen, Hilden, Germany) and genotyped using Sequenom's MassARRAY system (Sequenom, San Diego, California, USA).

2.3. Questionnaires

A structured face-to-face interview was conducted with each of the study subjects (cases and controls) by trained research assistants using a standard questionnaire (demographics, environmental exposures and family history of lung cancer/other cancer; and additional clinical characteristics for lung cancer for cases).

2.4. Statistical analysis

In the discovery stage, case-control comparisons of demographics and environment exposures were made separately between 103 EGFR mutant cases and matched controls as well as 78 EGFR wild-type cases and matched controls. Case-case comparisons were made between EGFR mutant and wild-type cases regarding clinical characteristics, presented as mean ± standard deviation (SD) or N (%) and compared by paired t-test or chi-square where appropriate. A p-value ≤0.05 was defined as statistically significant. Case-control comparisons of genotype frequencies were made between 103 EGFR mutant matched pairs and 78 EGFR wild-type matched pairs

separately to identify the genetic association in relation to EGFR mutant or wild-type lung cancer risk. It was explored in the following manner: (a) at the SNP level for individual SNP association using SNPstats [10]; (b) at the gene level for linkage disequilibrium (LD) by haplotype analysis using Haploview for those SNPs that showed an individual significant association [11]; and (c) at the pathway level for gene-gene and gene-environment interaction using a multifactor dimensionality reduction (MDR) model among significant/marginally significant SNPs as shown in the individual genetic association analysis [12]. The best genetic model was selected based on the lowest Akaike information criterion (AIC) value from the three models (additive, dominant, recessive) in SNPstats. Testing accuracy (TA) and cross-validation consistency (CVC) were used to choose the best model of MDR with permutation testing to determine statistical significance [12]. The Benjamini and Hochberg method was employed to control false discovery rate (FDR) [13]. SNPs with genotypes that departed significantly from Hardy-Weinberg equilibrium (HWE) or those with a call rate ≤ 90% were excluded from the association analysis.

In the validation stage, control-control comparisons of allelic frequencies were made between 103 Red Cross controls (matched controls for 103 EGFR mutant cases) and 103 CHB and 105 CHS public controls in 1000 genome. They are presented as N (%) and the analysis was performed by Chi-square. Case-case comparisons of demographics and environmental exposures were made between EGFR mutant cases recruited in the discovery stage (n=146) and EGFR mutant cases recruited in the validation stage (n=84). They are presented as mean ± standard deviation (SD) or N

(%) and the comparison was made using paired t-test or chi-square where appropriate. A p-value ≤0.05 was defined as statistically significant. Case-control comparisons were made between 84 EGFR mutant cases and 103 CHB public controls or 105 CHS public controls in 1000 genome separately to validate the genetic associations.

Sample size was calculated based on discordant pairs of matched cases and controls as shown in our pilot study by McNemar's Z-test. The proportion of discordant pairs of IL-6 rs 2069840 was 7.8% in matched cases and 18.4% in matched controls. Therefore, 89 EGFR mutant matched pairs were needed to detect an association of IL-6 rs2069840 (OR=3.62) with 80% power, 2-sided at 0.05 significant level. The proportion of discordant pairs of ATM rs611646 was 13.3% in matched cases and 37.8% in matched controls. Therefore, 65 EGFR wild-type matched pairs were needed to detect an association of ATM rs611646 (OR=2.97) with 80% power, 2-sided at 0.05 significant level.

3. Results

3.1 Discovery Stage

3.1.1 Case-control comparisons (103 EGFR mutant and 78 EGFR wild-type cases; matched controls): demographics and environmental exposures

Both EGFR mutant and wild-type matched pairs were balanced for age, gender and family history of lung or other cancers. Among 103 EGFR mutant and matched pairs,

significant differences were observed in environmental tobacco smoke (ETS) in the workplace (P<0.001, OR: 4.77, 95% CI: 2.26-10.08), textile in the workplace (P=0.03, OR: 4.44, 95% CI: 1.01-19.56), and chemical fumes in the workplace (P=0.002). Among 78 EGFR wild-type and matched pairs, a significant difference was observed only for chemical fumes in the workplace (P=0.008) (Table 1).

3.1.2 Case-Case comparisons (103 EGFR mutant cases vs. 78 EGFR wild-type cases): clinical characteristics

Advanced-stage lung adenocarcinoma was diagnosed in 72.2% of EGFR mutant and 78.6% of EGFR wild-type cases. Cough was the most common presenting feature (49.5% in EGFR mutant vs. 65.4% in EGFR wild-type). The major diagnostic method was bronchial biopsy (38.8% in EGFR mutant vs. 28.9% in EGFR wild-type). Anti-cancer therapy was received by 68% of EGFR mutant cases, of whom 70% were treated with a 1st generation EGFR tyrosine-kinase inhibitor (TKI) (erlotinib or gefitinib). Anti-cancer therapy was received by 62.8% of EGFR wild-type cases of whom 87.8% were either treated with systemic cytotoxic chemotherapy or entered into a clinical trial (Supplementary Table 2).

3.1.3 Case-control comparisons (103 EGFR mutant and 78 EGFR wild-type cases; matched controls): genetic associations

SNP Level

First, 51 SNPs in relation to EGFR mutant and wild-type lung adenocarcinoma were investigated individually. Six SNPs were significantly associated with EGFR mutant lung adenocarcinoma: rs238406 (P=0.028, OR=2.35, 95% CI: 1.18-4.87), rs238416 (P=0.038, OR=1.96, 95% CI: 1.03-3.71), rs1618536 (P=0.028, OR=2.35, 95% CI: 1.18-4.87) of ERCC2; rs2854508 (P=0.014, OR=2.32, 95% CI: 1.08-4.98), rs3213328 (P=0.006, OR=2.59, 95% CI: 1.29-5.20) of XRCC1; and rs2069840 (P=0.0059, OR=3.62, 95% CI: 1.37-9.52) of IL-6 (Table 2a). Two SNPs were significantly associated with EGFR wild-type lung adenocarcinoma: rs611624 (P=0.042, OR=2.04, 95% CI: 1.02-4.08) of ATM; rs50871 (P=0.0098, OR=0.23, 95% CI: 0.07-0.76) of ERCC2. Seven SNPs showed a marginally significant association with EGFR wild-type lung adenocarcinoma: rs189037 (P=0.056), rs599558 (P=0.06), rs228592 (P=0.06), rs609261 (P=0.08), rs227062 (P=0.09), rs664677 (P=0.10) and rs609429 (P=0.10) of ATM (Table 2b). These results suggest that SNPs that predispose to the development of EGFR mutant and wild-type lung adenocarcinoma might differ.

Gene Level

At a gene level, LD and haplotype analysis were performed separately in the EGFR mutant group (ERCC2, XRCC1 and IL-6) and EGFR wild-type group (ATM and ERCC2). Strong LD is shown with red blocks in Fig. 1. For the ATM gene, rs189037, rs599558, rs228592, rs609261, rs227062, rs664677 and rs609429 were in complete LD (D'=1) with rs611646. Haplotype analysis revealed that the haplotype GGG, the specific combination of genotypes in three SNPs (rs2069840, rs2069852 and rs2066992) of IL-6, was associated with increased risk of EGFR mutant lung adenocarcinoma

(Permutation P=0.02) (Fig. 1). The distribution of haplotype among EGFR mutant and wild-type cases and their respective matched controls is shown in the supplementary data (Supplementary Table 3 and 4).

Pathway Level

a) Gene-gene interaction

For EGFR mutant lung adenocarcinoma, one gene-gene interaction in the DNA repair pathway was identified in which ERCC2 rs238406 and XRCC1 rs3213328 (TA: 0.6117, CVC: 10/10, Permutation P=0.027) jointly increased the risk three-fold (OR: 3.15, 95% CI: 1.73-5.77) (Fig. 2A).

For EGFR wild-type lung adenocarcinoma, another gene-gene interaction in the DNA repair pathway was identified in which ERCC2 (rs50871) and ATM (rs611646 and rs599558) (TA: 0.6603, CVC: 10/10, Permutation P=0.006) jointly increased the risk 4.5-fold (OR=4.58, 95% CI: 2.33-9.04) (Fig. 2B).

b) Gene-environment interaction

For EGFR mutant lung adenocarcinoma, synergistic effects of ERCC2 rs238406 (DNA repair pathway) and ETS in the workplace (TA: 0.6262, CVC: 10/10, Permutation P=0.018) increased the risk 4-fold (OR= 4.08, 95% CI: 2.24-7.43) (Fig. 3A). When this synergistic effect of ERCC2/ETS in the workplace was combined with XRCC1 rs3213328 (TA: 0.6796, CVC: 10/10, Permutation P=0.0000-0.0001), the risk was further increased almost 6-fold (OR=5.92, 95% CI: 3.25-10.80) (Fig. 3B). Similarly, the

synergistic effects of IL-6 rs2069840 (inflammatory pathway) and ETS in the workplace (TA: 0.6068, CVC: 10/10, Permutation P=0.01) upgraded the risk to approximately 5-fold (OR: 4.69, 95% CI: 2.22-9.86) (Fig. 3C).

For EGFR wild-type lung adenocarcinoma, synergistic effects of ATM rs611646 (DNA repair pathway), ERCC2 rs50871 (DNA repair pathway) and ETS in the workplace (TA: 0.6154, CVC: 8/10, Permutation P=0.08) increased the risk to almost 5-fold (OR: 4.89, 95% CI: 2.43-9.86) (Fig. 3D).

3.2 Validation Stage

3.2.1 Control-control comparisons (103 Red Cross control vs. 103 CHB / 105CHS public controls in 1000 genome): allele frequencies

Firstly, allele frequencies of matched controls for 103 EGFR mutant cases were compared with 103 CHB and 105 CHS public controls in 1000 genome database for the 43 SNPs that were successfully genotyped (call rate >90%) and in accordance with HWE. The allele frequencies of all SNPs were similar to those of the public healthy controls, confirming that the controls in this study were presumably “healthy” (Supplementary Table 5).

3.2.2 Case-case comparisons (146 EGFR mutant cases recruited in discovery stage vs. 84 EGFR mutant cases recruited in validation stage): *demographics and environmental exposures*

Secondly, comparisons were made between EGFR mutant never-smoking lung adenocarcinoma cases recruited in the discovery stage and those in the validation stage. Demographics and environmental factors were comparable (Supplementary Table 6).

Lastly, two potentially functional SNPs (ERCC2 rs238406 and IL-6 rs2069840) indicated by F-SNP (http://compbio.cs.queensu.ca/F-SNP/) were selected for validation among 84 independent EGFR mutant cases and compared with 103 CHB and 105 CHS controls. The same genetic model to that used in the initial study was chosen in the validation [14].

3.2.3 *Case-control comparisons (84 EGFR mutant cases vs. 103 CHB controls; 84 EGFR mutant cases vs. 105 CHS controls): genetic associations*

The genotype frequencies of 84 cases were similar to those of the 103 cases in the discovery stage. The association of IL-6 rs2069840 in the validation stage was in the same direction with similar effect size to the discovery stage (P=0.02, OR: 2.76, 95% CI: 1.13-6.72) suggesting consistency (Table 3).

In comparison with CHS controls, the association of IL-6 rs2069840 was not significant. ERCC2 rs238406 showed a marginally significant association with EGFR mutant lung

adenocarcinoma in the same direction to that seen in the discovery stage (P=0.096) (Table 3).

4. Discussion

A large number of genetic studies have investigated the association of two promoter SNPs of IL-6 (rs1800795 (-174G/C) and rs1800796 (-572C/G, also known as-634G/C)) with risk of lung cancer. Nonetheless no conclusion can be drawn from meta-analysis even when restricted to the Chinese population. The inconsistent results might have been confounded by the heterogeneity in demographics, family history of cancer, environmental exposures (such as second hand smoke, cooking fumes) and different genotyping methods. For example, Bai [5] used Taqman genotyping, while Chen [6] used polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) in genotyping rs1800796. Lim [7] took into account family history of cancer and environmental exposure to tobacco smoke in data analysis, but these were not considered in the studies by Bai and Chen. In contrast, our current study tested SNPs rs2069840 (intron), rs2066992 (intron) and rs2069852 (coding region) of IL-6 by using MassARRAY, while family history of cancer and environmental factors were considered in multivariable analysis. Our findings reveal that differential driver mutations of lung cancer among never-smokers (i.e. EGFR mutant or wild-type) may further explain the previously reported discrepant genetic risks for lung cancer.

A recent meta-analysis reported that genetic variants in the ROS1/DCBLD1 gene (rs9387478) and HLA-DPB1 gene (rs2179920) were strongly associated with EGFR positive lung adenocarcinoma compared with EGFR negative cases (case-case comparison). It was the first study to report such a differential association by EGFR status, suggesting that specific germline variants might influence the acquisition of

specific mutational patterns in lung adenocarcinoma [15]. Nonetheless owing to the vast range of independent testing performed using GWAS, lack of consideration of heterogeneity in environmental and other host factors, as well as possible gene-environment interactions, the overall genetic effects were weak (OR <1.5) with relatively lower power.

In this two-stage case-control study, IL-6 rs2069840 and ERCC2 rs238406 were significantly/marginally significantly associated with the development of EGFR mutant lung adenocarcinoma in the validation stage, while eight SNPs of the ATM gene were significantly/marginally significantly associated with EGFR wild-type lung adenocarcinoma among never-smoking Chinese in the discovery stage. This initial study represents an important discovery in distinguishing SNPs that predispose an individual to EGFR mutant or wild-type lung adenocarcinoma, and attempts to replicate such association and provide fundamental clues to explore the functional role of these SNPs.

As one of the notable cytokines involved in the inflammation-to-cancer axis, IL-6 plays an essential role in lung carcinogenesis through several signalling pathways, primarily JAK/STAT3 [16-18]. STAT3 is an important signalling mediator in malignant disease and is persistently activated in 22 to 65% of NSCLC cases [19-21]. STAT3 is also involved in one of the EGFR downstream pathways [22]. The mechanism whereby EGFR mutant drives STAT3 activation is dependent on upregulation of IL-6 [23]. IL-6, acting in an EGFR-dependent (paracrine) and independent (autocrine) manner,

activates STAT3 leading to tumour growth and progression [23, 24]. Overexpression of IL-6 has been shown to be associated with decreased methylation of the EGFR promoter and enhanced EGFR protein expression, thereby contributing to the growth of cholangiocarcinoma [25]. IL6/JAK2/STATs pathway also upregulated DNA methyl transferase 1 and enhanced lung cancer stem cell proliferation [26]. In fact, the "crosstalk" between EGFR and IL-6 signalling pathway may contribute to the tumorigenesis of lung cancer (Supplementary Fig. 1). Song et al. illustrated that IL-6 antibody, siltuximab, could completely inhibit STAT3 tyrosine phosphorylation in NSCLC cells, and a combination of erlotinib and siltuximab could result in dual inhibition of lung cancer growth [27]. Yao et al. indicated that adjunctive therapies designed to either control inflammation and/or decrease the bioavailability of IL-6 may provide an effective means to improve response to EGFR TKI treatment in lung cancer [28].

The tumorigenic role of IL-6 in lung cancer can be biologically mediated through enhanced host susceptibility via SNPs. Interestingly, rs2069840 is located in the regulatory region of the IL-6 gene, suggesting a direct functional role in diseases mediated via IL-6. As predicted by F-SNP, rs2069840 may influence IL-6 protein level by binding with transcription factors. In this study, we established a novel finding that rs2069840 is an SNP that predisposes never-smokers to EGFR mutant lung adenocarcinoma.

Haplotype analysis is more powerful for mapping and characterizing disease-causing genes [29-31]. In this study, we revealed for the first time that haplotype GGG of

three SNPs (rs2069840, rs2069852, rs2066992) of IL-6 was significantly associated with the development of EGFR mutant lung adenocarcinoma. As predicted by F-SNP, rs2069840 and rs2066992 may be functional and may influence IL-6 protein level by binding transcriptional factors.

Moreover, a synergistic risk effect of IL-6 rs2069840 and ETS in the workplace for EGFR mutant lung adenocarcinoma was revealed, suggesting that rs2069840 and ETS might interact to confer risk in developing EGFR-mutated lung adenocarcinoma among never-smokers.

ERCC2 participates in the nucleotide excision repair (NER) pathway to remove DNA lesions [32]. rs238406 is located at the synonymous coding region, indicating a possible direct impact on ERCC2 protein structure. As predicted by the F-SNP database, it may affect ERCC2 protein level through an effect on mRNA splicing. The functional significance (FS) of rs238406 is 0.907 indicating a deleterious SNP [33].

Yin et al. revealed that ERCC2 rs238406 could confer susceptibility to lung adenocarcinoma among a never-smoking Chinese population [34]. In this study, ERCC2 rs238406 showed a marginally significant association with EGFR mutant lung adenocarcinoma. Nonetheless it is noteworthy that the genotype frequencies of CHB controls were quite different to those of CHS controls. Such “North-South discrepancy” has been reported in studies by Ling and Chen [35, 36]. As the genotype frequencies of

cases in our study were consistent across both the discovery and validation stages, the variations among controls might account for the negative replication.

ATM is involved in the double-stranded DNA breaks pathway for DNA repair. In line with our findings, Lo et al. reported a significant association of ATM SNPs with lung cancer risk among never-smokers [37]. This association was more evident in never-smokers with heavy ETS exposure and suggests an interaction between ATM SNPs and ETS in the carcinogenesis of lung cancer.

Some limitations of this study include: 1) the relatively small sample size; 2) possible selection bias and population stratification; 3) controls in the validation stage were not obtained from prospective sample collection and heterogeneity in CHB and CHS controls; 4) EGFR wild-type is a heterogeneous group; 5) biased allele frequency distribution of SNPs may also occur when sample size is small.

In conclusion, genetic susceptibility of never-smokers differs for EGFR mutant and wild-type lung adenocarcinoma. IL-6 rs2069840 is an important SNP that renders never-smoking Chinese in Hong Kong and Macau susceptible to EGFR-mutated lung adenocarcinoma.

Conflict of interest statement:

No authors of this manuscript have any financial or personal relationship with other people or organizations that could inappropriately influence the work. None of the authors have any conflicts of interest to declare.

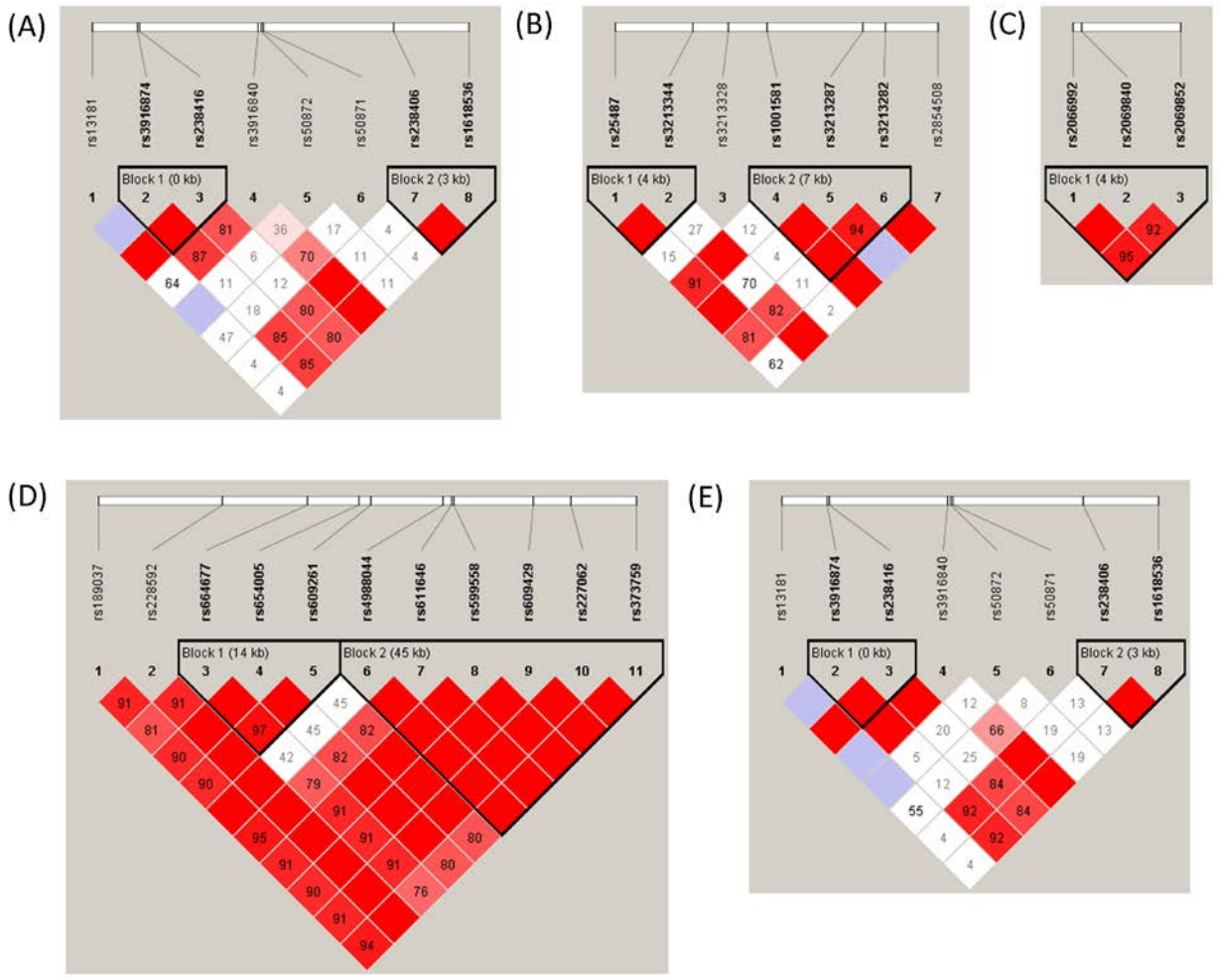

Figure 1. Linkage disequilibrium (LD) block and haplotype association in EGFR mutant and wild-type matched pairs by Haploview. LD plots were constructed by strong LD. The depth of red colour represents the computed pairwise D'. The number in the box is the D' value, the brightest red without number indicates complete LD (D'=1). (A) ERCC2 for EGFR mutant matched pairs. LD block 1: rs3916874 and rs238416; LD block 2: rs238406 and rs1618536. No statistical significance in haplotype analysis. (B) XRCC1 for EGFR mutant matched pairs LD block 1: rs25478 and rs3213344; LD block 2:rs1001581, rs3213287 and rs3213282. No statistical significance in haplotype analysis. (C) IL6 for EGFR mutant matched pairs. LD block 1: rs2066992, rs2069840

and rs2069852. Haplotype GGG was significantly associated with EGFR mutant lung ADC (Permutation P=0.02). (D) ATM for EGFR wild-type matched pairs. LD block 1: rs664677, rs654005, rs609261; LD block 2: rs4988044, rs611646, rs599558, rs609429, rs227062 and rs373759. No statistical significance in haplotype analysis. (E) ERCC2 for EGFR wild-type matched pairs. LD block 1:rs3916874 and rs238416; LD block 2: rs238406 and rs1618536. No statistical significance in haplotype analysis.

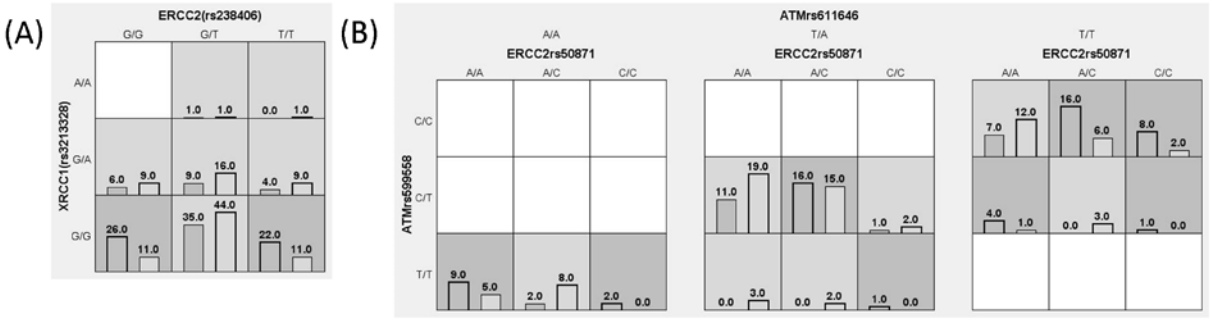

Figure 2. Gene-gene interaction for EGFR mutant and wild-type matched pairs by MDR. The graph shows the distribution of cases (left bars) and controls (right bars) for each combination of genotype/environmental factors. The high-risk cells are shaded dark grey while low-risk cells are shaded light grey according to the ratio of no. of cases vs. no. of controls >1 or ≤ 1. The white cells are empty cells. (A) Gene-gene interaction and joint effects of multiple genes in EGFR mutant lung adenocarcinoma. ERCC2 rs238406 and XRCC1 rs3213328: TA:0.6117, CVC:10/10, Permutation P=0.027, combined OR (95%CI): 3.15 (1.73-5.77). (B) Gene-gene interaction and joint effects of multiple genes in EGFR wild-type lung adenocarcinoma. ERCC2 rs50871, ATM rs611646, rs559558: TA:0.6603, CVC:10/10, Permutation P=0.006, combined OR (95%CI): 4.58 (2.33-9.04).

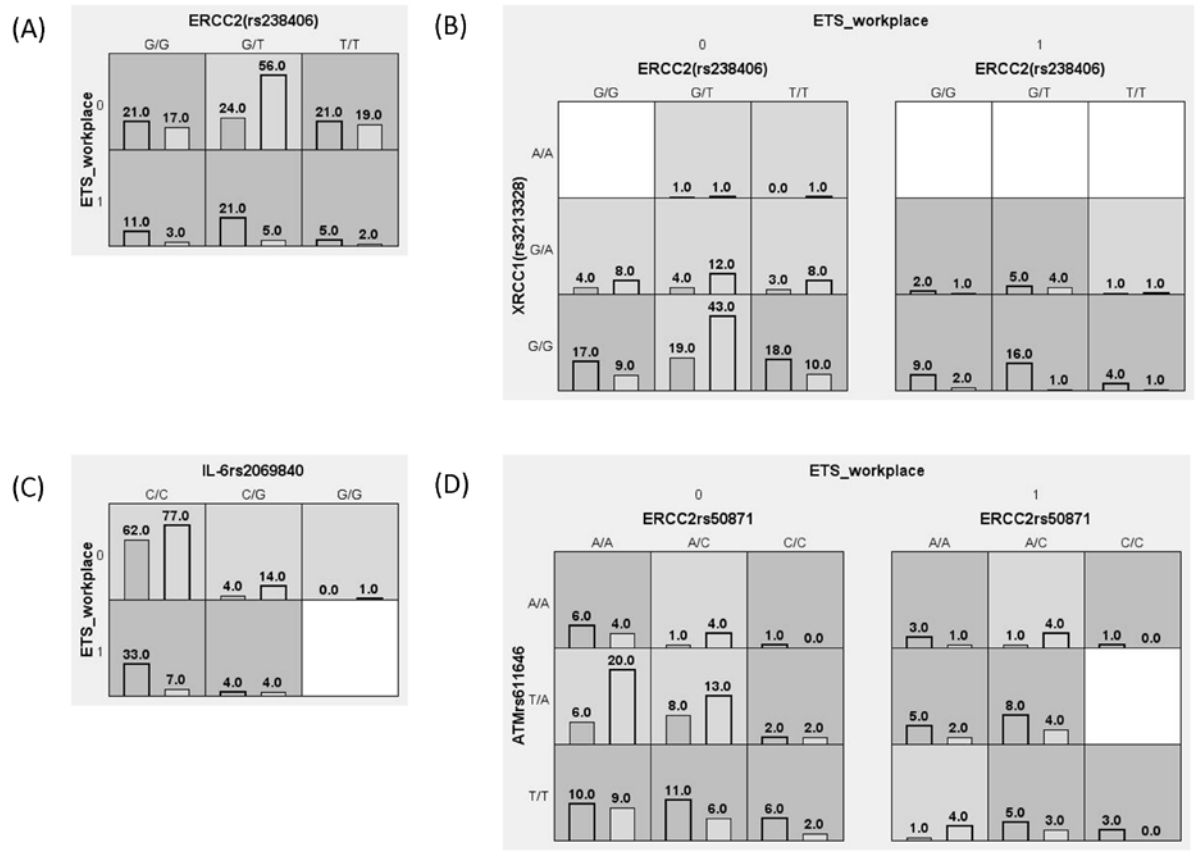

Figure 3. Gene-environment interaction of EGFR mutant and wild-type matched pairs by MDR. (A) Gene-environment interaction and joint effects of gene and environmental tobacco smoke in EGFR mutant lung adenocarcinoma (DNA repair pathway, 2 factors). ERCC2 rs238406, ETS in the workplace: TA:0.6262, CVC:10/10, Permutation P=0.018, combined OR (95%CI): 4.08 (2.24-7.43). (B) Gene-environment interaction and joint effects of gene and environmental tobacco smoke in EGFR mutant lung adenocarcinoma (DNA repair pathway, 3 factors). ERCC2 rs238406, XRCC1 rs3213328 and ETS in the workplace: TA:0.6796, CVC:10/10, Permutation P=0.000-0.0001, combined OR (95%CI): 5.92 (3.25-10.80). (C) Gene-environment interaction and joint effects of gene and environmental tobacco smoke in EGFR mutant lung

adenocarcinoma (Inflammatory pathway). IL6 rs2069840, ETS in the workplace: TA:0.6068, CVC:10/10, Permutation P=0.01, combined OR(95%CI): 4.69 (2.22-9.86). (D) Gene-environment interaction and joint effects of gene and environmental tobacco smoke in EGFR wild-type lung adenocarcinoma (DNA repair pathway). ERCC2 rs50871, ATM rs611646 and ETS in the workplace: TA:0.6154, CVC:8/10, Permutation P=0.08, combined OR (95%CI): 4.89 (2.43-9.86).

Supplementary figure 1. Possible crosstalk between EGFR and IL6 signalling pathway. PI3K/AKT/mTOR, RAS/RAF/MEK/ERK and JAK/STAT3 are three major EGFR downstream pathways. The mechanism whereby mutant EGFR drives STAT3 activation is dependent on the upregulation of IL6. IL6, acting in an EGFR-dependent (paracrine) and independent (autocrine) manner, activates STAT3. Overexpression of IL6 inhibits EGFR promotor methylation and increases EGFR protein expression.

Acknowledgements:

We thank all the participants in this study.

REFERENCES:

1. Nakamura H, Saji H. Worldwide trend of increasing primary adenocarcinoma of the lung. Surg Today 2014;44: 1004-1012.
2. Subramanian J, Govindan R. Lung cancer in never smokers: a review. J Clin Oncol 2007;25: 561-570.
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer 2007;7: 778-790.
4. Sun Y, Ren Y, Fang Z, Li C, Fang R, Gao B, Han X, Tian W, Pao W, Chen H, Ji H. Lung adenocarcinoma from East Asian never-smokers is a disease largely defined by targetable oncogenic mutant kinases. J Clin Oncol 2010;28: 4616-4620.
5. Bai L, Yu H, Wang H, Su H, Zhao J, Zhao Y. Genetic single-nucleotide polymorphisms of inflammation-related factors associated with risk of lung cancer. Med Oncol 2013;30: 414.
6. Chen J, Liu RY, Yang L, Zhao J, Zhao X, Lu D, Yi N, Han B, Chen XF, Zhang K, He J, Lei Z, Zhou Y, Pasche B, Li X, Zhang HT. A two-SNP IL-6 promoter haplotype is associated with increased lung cancer risk. J Cancer Res Clin Oncol 2013;139: 231-242.
7. Lim WY, Chen Y, Ali SM, Chuah KL, Eng P, Leong SS, Lim E, Lim TK, Ng AW, Poh WT, Tee A, Teh M, Salim A, Seow A. Polymorphisms in inflammatory pathway genes, host factors and lung cancer risk in Chinese female never-smokers. Carcinogenesis 2011;32: 522-529.
8. Kiyohara C, Horiuchi T, Takayama K, Nakanishi Y. Genetic polymorphisms involved in the inflammatory response and lung cancer risk: a case-control study in Japan. Cytokine 2014;65: 88-94.
9. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 2015;526: 68-74.
10. Sole X, Guino E, Valls J, Iniesta R, Moreno V. SNPStats: a web tool for the analysis of association studies. Bioinformatics 2006;22: 1928-1929.
11. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21: 263-265.
12. Motsinger AA, Ritchie MD. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Hum Genomics 2006;2: 318-328.
13. Reiner-Benaim A. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. Biom J 2007;49: 107-126.
14. Konig IR. Validation in genetic association studies. Brief Bioinform 2011;12: 253-258.
15. Seow WJ, Matsuo K, Hsiung CA, Shiraishi K, Song M, Kim HN, Wong MP, Hong YC, Hosgood HD, 3rd, Wang Z, Chang IS, Wang JC, Chatterjee N, Tucker M, Wei H, Mitsudomi T, Zheng W, Kim JH, Zhou B, Caporaso NE, Albanes D, Shin MH, Chung LP, An SJ, Wang P, Zheng H, Yatabe Y, Zhang XC, Kim YT, Shu XO, Kim YC, Bassig BA, Chang J, Ho JC, Ji BT, Kubo M, Daigo Y, Ito H, Momozawa Y, Ashikawa K, Kamatani Y, Honda T, Sakamoto H, Kunitoh H, Tsuta K, Watanabe SI, Nokihara H, Miyagi Y, Nakayama H, Matsumoto S, Tsuboi M, Goto K, Yin Z, Shi J, Takahashi A, Goto A, Minamiya Y, Shimizu K, Tanaka K, Wu T, Wei F, Wong JY, Matsuda F, Su J, Kim YH, Oh IJ, Song F, Lee VH, Su WC, Chen YM, Chang GC, Chen KY, Huang MS, Yang PC, Lin HC, Xiang YB, Seow A, Park JY, Kweon SS, Chen CJ, Li H, Gao YT, Wu C, Qian B, Lu D, Liu J, Jeon HS, Hsiao CF, Sung JS, Tsai YH, Jung YJ, Guo H, Hu Z, Wang WC, Chung CC, Lawrence C, Burdett L, Yeager M, Jacobs KB, Hutchinson A, et al. Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. Hum Mol Genet 2017;26: 454-465.

16. Ochoa CE, Mirabolfathinejad SG, Ruiz VA, Evans SE, Gagea M, Evans CM, Dickey BF, Moghaddam SJ. Interleukin 6, but not T helper 2 cytokines, promotes lung carcinogenesis. Cancer Prev Res (Phila) 2011;4: 51-64.
17. Lin WW, Karin M. A cytokine-mediated link between innate immunity, inflammation, and cancer. J Clin Invest 2007;117: 1175-1183.
18. Huang WL, Yeh HH, Lin CC, Lai WW, Chang JY, Chang WT, Su WC. Signal transducer and activator of transcription 3 activation up-regulates interleukin-6 autocrine production: a biochemical and genetic study of established cancer cell lines and clinical isolated human cancer cells. Mol Cancer 2010;9: 309.
19. Zimmer S, Kahl P, Buhl TM, Steiner S, Wardelmann E, Merkelbach-Bruse S, Buettner R, Heukamp LC. Epidermal growth factor receptor mutations in non-small cell lung cancer influence downstream Akt, MAPK and Stat3 signaling. J Cancer Res Clin Oncol 2009;135: 723-730.
20. Looyenga BD, Hutchings D, Cherni I, Kingsley C, Weiss GJ, Mackeigan JP. STAT3 is activated by JAK2 independent of key oncogenic driver mutations in non-small cell lung carcinoma. PLoS One 2012;7: e30820.
21. Jiang R, Jin Z, Liu Z, Sun L, Wang L, Li K. Correlation of activated STAT3 expression with clinicopathologic features in lung adenocarcinoma and squamous cell carcinoma. Mol Diagn Ther 2011;15: 347-352.
22. Zhong Z, Wen Z, Darnell JE, Jr. Stat3: a STAT family member activated by tyrosine phosphorylation in response to epidermal growth factor and interleukin-6. Science 1994;264: 95-98.
23. Gao SP, Mark KG, Leslie K, Pao W, Motoi N, Gerald WL, Travis WD, Bornmann W, Veach D, Clarkson B, Bromberg JF. Mutations in the EGFR kinase domain mediate STAT3 activation via IL-6 production in human lung adenocarcinomas. J Clin Invest 2007;117: 3846-3856.
24. Grivennikov S, Karin M. Autocrine IL-6 signaling: a key event in tumorigenesis? Cancer Cell 2008;13: 7-9.
25. Wehbe H, Henson R, Meng F, Mize-Berge J, Patel T. Interleukin-6 contributes to growth in cholangiocarcinoma cells by aberrant promoter methylation and gene expression. Cancer Res 2006;66: 10517-10524.
26. Liu CC, Lin JH, Hsu TW, Su K, Li AF, Hsu HS, Hung SC. IL-6 enriched lung cancer stem-like cell population by inhibition of cell cycle regulators via DNMT1 upregulation. Int J Cancer 2015;136: 547-559.
27. Song L, Rawal B, Nemeth JA, Haura EB. JAK1 activates STAT3 activity in non-small-cell lung cancer cells and IL-6 neutralizing antibodies can suppress JAK1-STAT3 signaling. Mol Cancer Ther 2011;10: 481-494.
28. Yao Z, Fenoglio S, Gao DC, Camiolo M, Stiles B, Lindsted T, Schlederer M, Johns C, Altorki N, Mittal V, Kenner L, Sordella R. TGF-beta IL-6 axis mediates selective and adaptive mechanisms of resistance to molecular targeted therapy in lung cancer. Proc Natl Acad Sci U S A 2010;107: 15535-15540.
29. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 2001;9: 291-300.
30. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29: 233-237.
31. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118: 1590-1605.

32. Leadon SA, Cooper PK. Preferential repair of ionizing radiation-induced damage in the transcribed strand of an active human gene is defective in Cockayne syndrome. Proc Natl Acad Sci U S A 1993;90: 10499-10503.
33. Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 2008;36: D820-824.
34. Yin J, Li J, Ma Y, Guo L, Wang H, Vogel U. The DNA repair gene ERCC2/XPD polymorphism Arg 156Arg (A22541C) and risk of lung cancer in a Chinese population. Cancer Lett 2005;223: 219-226.
35. Ling Y, Jin Z, Su M, Zhong J, Zhao Y, Yu J, Wu J, Xiao J. VCGDB: a dynamic genome database of the Chinese population. BMC Genomics 2014;15: 265.
36. Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, Liu J. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet 2009;85: 775-785.
37. Lo YL, Hsiao CF, Jou YS, Chang GC, Tsai YH, Su WC, Chen YM, Huang MS, Chen HL, Yang PC, Chen CJ, Hsiung CA. ATM polymorphisms and risk of lung cancer among never smokers. Lung Cancer 2010;69: 148-154.

Table 1. Baseline demographics and environmental exposure of 103 EGFR mutant and 78 EGFR wild-type cases and their respective age- and gender- matched controls. [a]

	103 EGFR mutant matched pairs N (%)			78 EGFR wild-type matched pairs N (%)		
	Case (n=103)	**Control (n=103)**	**P OR (95%CI)**	**Case (n=78)**	**Control (n=78)**	**P OR (95%CI)**
Age, years (Mean± SD)	58.1±7.5	57.5± 5.0	0.14	54.5±12.9	54.2± 7.6	0.65
Gender						
Male	29 (28.2)	29 (28.2)	1.00	32 (41.0)	32 (41.0)	1.00
Female	74 (71.8)	74 (71.8)		46 (59.0)	46 (59.0)	

Family history of lung cancer	15 (14.6)	17 (17.2)		0.61	11 (14.5)	14 (17.9)	0.56
Family history of other cancers	39 (37.9)	40 (41.2)		0.63	31 (40.8)	26 (33.3)	0.34
ETS exposure in the home	38 (38.4) (n=99)	27 (27.3) (n=99)	0.10		28 (36.4) (n=77)	21 (26.9) (n=78)	0.21
ETS exposure in the workplace	37 (37.4) (n=99)	11 (11.1) (n=99)	**<0.001 4.77 (2.26-10.08)**		27 (35.5) (n=76)	19 (24.4) (n=78)	0.13
Cooking fumes in the home	28 (68.3) (n=41)	60 (60.6) (n=99)	0.39		32 (65.3) (n=49)	41 (52.6) (n=78)	0.16
Cooking fumes in the workplace	5 (12.2) (n=41)	4 (4.0) (n=99)	0.07		6 (12.2) (n=49)	6 (7.7) (n=78)	0.39

Asbestos dust in the workplace	0 (0) (n=41)	0 (0) (n=99)	N.A	0 (0) (n=49)	0 (0) (n=78)	N.A
Silica dust in the workplace	1 (2.4) (n=41)	1 (1.0) (n=99)	0.52	1 (2.0) (n=49)	2 (2.6) (n=78)	0.85
Wood dust in the workplace	1 (2.4) (n=41)	0 (0) (n=99)	0.12	0 (0) (n=49)	6 (6.4) (n=78)	0.07
Coal dust in the workplace	0 (0) (n=41)	0 (0) (n=99)	N.A	0 (0) (n=49)	0 (0) (n=78)	N.A
Textile in the workplace	5 (12.2) (n=41)	3 (3.0) (n=99)	**0.03** **4.44 (1.01-19.56)**	3 (6.1) (n=49)	4 (5.1) (n=78)	0.81
Chemical fumes in the workplace	4 (9.8) (n=41)	0 (0) (n=99)	**0.002** **N.A**	5 (10.2) (n=49)	0 (0) (n=78)	**0.008** **N.A**

Burning fumes in the workplace	1 (2.4) (n=41)	1 (1.0) (n=99)	0.52	3 (6.1) (n=49)	3 (3.8) (n=78)	0.56
Radioactive material in the workplace	0 (0) (n=41)	0 (0) (n=99)	N.A	0 (0) (n=49)	0 (0) (n=78)	N.A

ETS, Environmental Tobacco Smoke; N. A, not applicable.
[a] n: Number included in analysis. All the cases and controls were suitable for analysis of ETS exposure. Cases recruited after May 2013 adopting new questionnaire and all controls were suitable for analysis of other environmental exposures.

Table 2a. Individual SNP analysis for 103 pairs of EGFR mutant never-smoking lung adenocarcinoma and matched controls.

Gene	SNP [d]	N (%)			P [a]	P-BH [b]
		genotype	103 EGFR mutant never-smoking adenocarcinoma	103 matched control	OR(95%CI) genetic model [c]	
ERCC2	rs238406	G/G	32 (31.1)	20 (19.6)	**0.028**	0.24
		G/T	45 (43.7)	61 (59.8)	**2.35 (1.13-4.87)**	
		T/T	26 (25.2)	21 (20.6)	Additive (G/T vs. G/G)	
ERCC2	rs3916840	G/G	85 (82.5)	85 (83.3)	0.13	0.45
		G/A	16 (15.5)	17 (16.7)	NA	
		A/A	2 (1.9)	0 (0)	Recessive (G/G vs. G/A+A/A)	
ERCC2	rs1618536	C/C	32 (31.1)	20 (19.6)	**0.028**	0.24
		T/C	45 (43.7)	61 (59.8)	**2.35 (1.13-4.87)**	
		T/T	26 (25.2)	21 (20.6)	Additive (T/C vs. C/C)	
ERCC2	rs13181	T/T	82 (79.6)	83 (81.4)	0.77	0.84
		G/T	19 (18.4)	17 (16.7)	0.74 (0.10-5.54)	
		G/G	2 (1.9)	2 (2.0)	Recessive (G/G vs. G/T+T/T)	
ERCC2	rs238416	C/C	38 (36.9)	25 (24.3)	**0.038**	0.27
		C/T	42 (40.8)	53 (51.5)	**1.96 (1.03-3.71)**	
		T/T	23 (22.3)	25 (24.3)	Dominant (C/T+T/T vs. C/C)	
ERCC2	rs50871	A/A	50 (48.5)	53 (51.5)	0.56	0.67
		C/A	44 (42.7)	40 (38.8)	1.36 (0.49-3.77)	
		C/C	9 (8.7)	10 (9.7)	Recessive (C/C vs. C/A+A/A)	
ERCC2	rs3916874	C/C	55 (53.4)	64 (62.1)	0.22	0.45
		C/G	43 (41.8)	33 (32)	0.69 (0.38-1.25)	

		G/G	5 (4.8)	6 (5.8)	Dominant (G/G +C/G vs. C/C)	
ERCC2	rs50872	G/G	59 (57.3)	70 (68)	0.32	0.46
		G/A	41 (39.8)	31 (30.1)	0.74 (0.40-1.35)	
		A/A	3 (2.9)	2 (1.9)	Dominant (G/A+ A/A vs. G/G)	
XRCC1	rs3213282	C/C	62 (60.2)	53 (52.0)	0.10	0.45
		G/C	36 (35.0)	44 (43.1)	1.64 (0.90-2.99)	
		G/G	5 (4.8)	5 (4.9)	Dominant (G/C+G/G vs. C/C)	
XRCC1	rs3213287	T/T	77 (74.8)	77 (76.2)	0.3	0.46
		C/T	25 (24.3)	20 (19.8)	2.94 (0.31-27.66)	
		C/C	1 (0.9)	4 (4.0)	Recessive (C/C vs. C/T+T/T)	
XRCC1	rs2854508	T/T	86 (83.5)	75 (73.5)	**0.014**	0.20
		A/T	15 (14.6)	27 (26.5)	**2.32 (1.08-4.98)**	
		A/A	2 (1.9)	0 (0)	Additive (A/T vs. T/T)	
XRCC1	rs3213344	G/G	49 (47.6)	56 (54.9)	0.13	0.45
		G/C	45 (43.7)	42 (41.2)	0.39 (0.11-1.39)	
		C/C	9 (8.7)	4 (3.9)	Recessive (C/C vs. G/G+G/C)	
XRCC1	rs25487	C/C	57 (55.3)	53 (52.5)	0.67	0.76
		T/C	38 (36.9)	39 (38.6)	1.14 (0.63-2.06)	
		T/T	8 (7.8)	9 (8.9)	Dominant (T/C+T/T vs. C/C)	
XRCC1	rs1001581	C/C	42 (40.8)	40 (38.8)	0.66	0.76
		C/T	48 (46.6)	47 (45.6)	1.21 (0.52-2.81)	
		T/T	13 (12.6)	16 (15.5)	Recessive (T/T vs. C/T+C/C)	
XRCC1	rs3213328	G/G	83 (80.6)	67 (65.0)	**0.006**	0.13
		A/G	19 (18.4)	34 (33.0)	**2.59 (1.29-5.20)**	
		A/A	1 (1.0)	2 (1.9)	Dominant (A/G+A/A vs. G/G)	
ATM	rs599558	C/C	27 (26.2)	29 (28.4)	0.21	0.45
		C/T	53 (51.5)	57 (55.9)	0.62 (0.29-1.31)	
		T/T	23 (22.3)	16 (15.7)	Recessive (T/T vs. C/T+C/C)	
ATM	rs227062	G/G	27 (26.2)	29 (28.7)	0.22	0.45

		G/A	53 (51.5)	56 (55.5)	0.63 (0.30-1.33)	
		A/A	23 (22.3)	16 (15.8)	Recessive (A/A vs. G/A+G/G)	
ATM	rs4988044	T/T	88 (85.4)	89 (87.2)	0.99	0.99
		C/T	15 (14.6)	13 (12.8)	1.00 (0.42-2.40)	
		C/C	0 (0)	0 (0)	C/T vs. T/T	
ATM	rs664677	C/C	27 (26.2)	29 (28.4)	0.27	0.46
		C/T	56 (54.4)	59 (57.8)	0.64 (0.29-1.41)	
		T/T	20 (19.4)	14 (13.7)	Recessive (T/T vs. C/T+C/C)	
ATM	rs189037	G/G	26 (25.5)	34 (33.7)	0.18	0.45
		G/A	52 (51.0)	51 (50.5)	0.60 (0.28-1.26)	
		A/A	24 (23.5)	16 (15.8)	Recessive (A/A vs. G/G+G/A)	
ATM	rs654005	G/G	30 (29.1)	33 (32.7)	0.32	0.46
		A/G	55 (53.4)	55 (54.5)	0.66 (0.29-1.50)	
		A/A	18 (17.5)	13 (12.9)	Recessive (A/A vs. A/G+G/G)	
ATM	rs609429	G/G	27 (26.2)	29 (28.7)	0.22	0.45
		G/C	53 (51.5)	56 (55.5)	0.63 (0.30-1.33)	
		C/C	23 (22.3)	16 (15.8)	Recessive (C/C vs. G/C+G/G)	
ATM	rs609261	T/T	30 (29.1)	35 (34.0)	0.29	0.46
		T/C	55 (53.4)	55 (53.4)	0.64 (0.28-1.45)	
		C/C	18 (17.5)	13 (12.6)	Recessive (C/C vs. T/T+T/C)	
ATM	rs228592	C/C	27 (26.2)	30 (29.1)	0.19	0.45
		C/A	53 (51.5)	57 (55.3)	0.61 (0.29-1.28)	
		A/A	23 (22.3)	16 (15.5)	Recessive (A/A vs. A/C +C/C)	
ATM	rs373759	C/C	41 (39.8)	47 (45.6)	0.13	0.45
		T/C	46 (44.7)	47 (45.6)	0.50 (0.20-1.24)	
		T/T	16 (15.5)	9 (8.7)	Recessive (T/T vs. T/C+C/C)	
ATM	rs611646	T/T	31 (30.1)	39 (37.9)	0.17	0.45
		T/A	52 (50.5)	51 (49.5)	0.57 (0.26-1.28)	
		A/A	20 (19.4)	13 (12.6)	Recessive (A/A vs. T/T+T/A)	

OGG1	rs293795	A/A	93 (90.3)	87 (85.3)	0.24	0.45
		G/A	10 (9.7)	14 (13.7)	1.72 (0.69-4.28)	
		G/G	0 (0)	1 (1.0)	Dominant (G/G+G/A vs. A/A)	
OGG1	rs1052133	G/G	37 (35.9)	31 (30.1)	0.19	0.45
		G/C	51 (49.5)	50 (48.5)	1.66 (0.77-3.58)	
		C/C	15 (14.6)	22 (21.4)	Recessive (C/C vs. G/G+G/C)	
OGG1	rs2072668	G/G	38 (36.9)	31 (30.1)	0.19	0.45
		G/C	50 (48.5)	50 (48.5)	1.66 (0.77-3.58)	
		C/C	15 (14.6)	22 (21.4)	Recessive (C/C vs. G/G+G/C)	
MLH1	rs3172297	T/T	92 (89.3)	90 (88.2)	0.069	0.42
		T/C	11 (10.7)	11 (10.8)	NA	
		C/C	0 (0)	1 (1.0)	Recessive (C/C vs. T/C+T/T)	
MLH1	rs1800734	A/A	34 (33.0)	37 (35.9)	0.11	0.45
		G/A	56 (54.4)	45 (43.7)	1.89 (0.85-4.20)	
		G/G	13 (12.6)	21 (20.4)	Recessive (G/G vs. G/A+A/A)	
IL-6	rs2069852	A/A	42 (40.8)	43 (42.2)	0.24	0.45
		G/A	52 (50.5)	46 (45.1)	1.78 (0.67-4.74)	
		G/G	9 (8.7)	13 (12.8)	Recessive (G/G vs. G/A+A/A)	
IL-6	rs2069840	C/C	95 (92.2)	84 (81.5)	**0.0059**	0.13
		C/G	8 (7.8)	18 (17.5)	**3.62 (1.37-9.52)**	
		G/G	0 (0)	1 (1)	Dominant (G/G+C/G vs. C/C)	
IL-6	rs2066992	T/T	55 (53.4)	54 (52.4)	0.47	0.59
		G/T	45 (43.7)	43 (41.8)	1.71 (0.39-7.47)	
		G/G	3 (2.9)	6 (5.8)	Recessive (G/G vs. G/T+T/T)	
IL-10	rs3024490	A/A	55 (53.4)	52 (51)	0.80	0.84
		C/A	40 (38.8)	42 (41.2)	1.16 (0.38-3.51)	
		C/C	8 (7.8)	8 (7.8)	Recessive (C/C vs. C/A+A/A)	
IL-10	rs1800871	A/A	55 (53.4)	52 (51.0)	0.80	0.84
		A/G	40 (38.8)	42 (41.2)	1.16 (0.38-3.51)	

		G/G	8 (7.8)	8 (7.8)	Recessive (G/G vs. A/G+ A/A)	
GPC5	rs2352028	C/C	65 (63.7)	56 (54.9)	0.26	0.46
		C/T	35 (34.3)	42 (41.2)	1.41 (0.78-2.56)	
		T/T	2 (2.0)	4 (3.9)	Dominant (C/T+ T/T vs. C/C)	
CY1A1	rs4646903	A/A	35 (34.0)	28 (27.4)	0.35	0.47
		A/G	46 (44.7)	58 (56.9)	0.70 (0.33-1.49)	
		G/G	22 (21.4)	16 (15.7)	Recessive (G/G vs. A/A+A/G)	
CY1A1	rs4646422	C/C	81 (78.6)	78 (75.7)	0.39	0.51
		C/T	19 (18.4)	23 (22.3)	1.36 (0.67-2.76)	
		T/T	3 (2.9)	2 (1.9)	Dominant (C/T+T/T vs. C/C)	
CY1A1	rs4646421	G/G	36 (35.0)	28 (27.7)	0.34	0.47
		A/G	45 (43.7)	57 (56.4)	0.69 (0.32-1.48)	
		A/A	22 (21.4)	16 (15.8)	Recessive (A/A vs. A/G+G/G)	
CLPTMIL	rs401681	C/C	46 (44.7)	47 (46.1)	0.85	0.87
		T/C	48 (46.6)	47 (46.1)	0.94 (0.52-1.70)	
		T/T	9 (8.7)	8 (7.8)	Dominant (T/C+T/T vs. C/C)	
CLPTMIL	rs402710	C/C	44 (42.7)	47 (46.1)	0.53	0.65
		T/C	50 (48.5)	47 (46.1)	0.83 (0.46-1.49)	
		T/T	9 (8.7)	8 (7.8)	Dominant (T/T+T/C vs. C/C)	
C3of21	rs2131877	A/A	30 (29.1)	30 (29.4)	0.31	0.46
		A/G	52 (50.5)	56 (54.9)	0.67 (0.32-1.44)	
		G/G	21 (20.4)	16 (15.7)	Recessive (G/G vs. A/A+A/G)	

[a] P value adjusted for age, gender, family history of lung cancer, family history of other cancer, environmental exposure to smoke in the home, environmental exposure to smoke in the workplace.

[b] P-BH adjusted for FDR-BH.

[c] The genetic model included recessive, dominant, additive.

[d] XRCC1 rs1799778; XPC rs1106087, rs3731055, rs2279017; ATM rs664982 were excluded because the call rate<90%; MDM 2 rs2279744; TP63 rs10937405; IL10 rs1878672 were excluded because controls departed from HWE.

Table 2b. Individual SNP analysis in 78 pairs of EGFR wild-type never-smoking lung adenocarcinoma and matched controls.

Gene	SNP [d]	N (%)			P [a]	P-BH[b]
		genotype	78 EGFR wild-type never-smoking adenocarcinoma	78 matched control	OR(95%CI) genetic model [c]	
ERCC2	rs238406	G/G	14 (17.9)	20 (25.6)	0.31	0.58
		G/T	45 (57.7)	44 (56.4)	0.66 (0.29-1.49)	
		T/T	19 (24.4)	14 (17.9)	Dominant (G/T+T/T vs. G/G)	
ERCC2	rs3916840	G/G	67 (85.9)	64 (82.1)	1.00	1
		G/A	11 (14.1)	14 (17.9)	1.52 (0.62-3.75)	
		A/A	0 (0)	0 (0)	G/A vs. G/G	
ERCC2	rs1618536	C/C	14 (17.9)	20 (25.6)	0.31	0.58
		C/T	45 (57.7)	44 (56.4%)	0.66 (0.29-1.49)	
		T/T	19 (24.4)	14 (17.9%)	Dominant (C/T+T/T vs. C/C)	
ERCC2	rs13181	T/T	61 (78.2)	64 (82.1)	0.25	0.58
		G/T	16 (20.5)	14 (17.9)	NA	
		G/G	1 (1.3)	0 (0)	Recessive (G/G vs. G/T+T/T)	
ERCC2	rs238416	C/C	15 (19.2)	21 (26.9)	0.28	0.58
		C/T	47 (60.3)	45 (57.7)	0.65 (0.30-1.42)	
		T/T	16 (20.5)	12 (15.4)	Dominant (C/T+T/T vs. C/C)	
ERCC2	rs50871	A/A	31 (39.7)	40 (51.3)	**0.0098**	0.40
		A/C	34 (43.6)	34 (43.6)	**0.23 (0.07-0.76)**	
		C/C	13 (16.7)	4 (5.1)	Recessive (C/C vs. A/C+A/A)	
ERCC2	rs3916874	C/C	43 (55.1)	40 (51.3)	0.32	0.58
		C/G	34 (43.6)	35 (44.9)	2.99 (0.29-30.37)	
		G/G	1 (1.3)	3 (3.8)	Recessive (G/G vs. C/G+C/C)	

ERCC2	rs50872	G/G	48 (61.5)	50 (64.1)	0.60	0.71
		G/A	28 (35.9)	26 (33.3)	0.83 (0.42-1.65)	
		A/A	2 (2.6)	2 (2.6)	Dominant (G/A+A/A vs. G/G)	
XRCC1	rs3213282	C/C	42 (53.9)	45 (57.7)	0.70	0.79
		G/C	32 (41.0)	28 (35.9)	0.88 (0.46-1.69))	
		G/G	4 (5.1)	5 (6.4)	Dominant (G/C+ G/G vs. C/C)	
XRCC1	rs3213287	T/T	63 (80.8)	57 (75)	0.29	0.58
		T/C	14 (17.9)	17 (22.4)	1.53 (0.69-3.37)	
		C/C	1 (1.3)	2 (2.6)	Dominant (T/C+C/C vs. T/T)	
XRCC1	rs2854508	T/T	54 (69.2)	61 (78.2)	0.15	0.47
		A/T	22 (28.2)	16 (20.5)	0.58 (0.27-1.22)	
		A/A	2 (2.6)	1 (1.3)	Dominant (A/T+ T/T vs. A/A)	
XRCC1	rs3213344	G/G	45 (57.7)	37 (47.4)	0.07	0.40
		G/C	29 (37.2)	30 (38.5)	2.92 (0.85-10.04)	
		C/C	4 (5.1)	11 (14.1)	Recessive (G/G vs. G/C+C/C)	
XRCC1	rs25487	C/C	43 (55.1)	44 (58.7)	0.31	0.58
		C/T	33 (42.3)	27 (36.0)	2.46 (0.41-14.80)	
		T/T	2 (2.6)	4 (5.3)	Recessive (T/T vs. C/T+C/C)	
XRCC1	rs1001581	C/C	32 (41.0)	28 (35.9)	0.51	0.69
		C/T	38 (48.7)	42 (53.9)	1.25 (0.64-2.47)	
		T/T	8 (10.3)	8 (10.3)	Dominant (T/T+C/T vs. C/C)	
ATM	rs599558	C/C	31 (39.7)	20 (25.6)	0.06	0.40
		C/T	33 (42.3)	40 (51.3)	1.96 (0.96-4.02)	
		T/T	14 (17.9)	18 (23.1)	Dominant (C/T+T/T vs. C/C)	
ATM	rs227062	G/G	31 (39.7)	20 (26.3)	0.09	0.40
		G/A	33 (42.3)	38 (50.0)	1.87 (0.91-3.84)	
		A/A	14 (17.9)	18 (23.7)	Dominant (G/A+A/A vs. G/G)	
ATM	rs4988044	T/T	63 (80.8)	64 (82.1)	0.63	0.73
		C/T	15 (19.2)	14 (17.9)	0.81 (0.34-1.91)	

		C/C	0 (0)	0 (0)	C/T vs. T/T	
ATM	rs664677	C/C	32 (41.0)	22 (28.6)	0.10	0.40
		C/T	35 (44.9)	40 (52.0)	1.80 (0.89-3.62)	
		T/T	11 (14.1)	15 (19.5)	Dominant (C/T+T/T vs. C/C)	
ATM	rs189037	G/G	30 (39.5)	20 (26)	0.056	0.40
		G/A	31 (40.8)	41 (53.2)	2.03 (0.98-4.23)	
		A/A	15 (19.7)	16 (20.8)	Dominant (G/A +A/A vs. G/G)	
ATM	rs654005	G/G	34 (43.6)	23 (30.3)	0.11	0.40
		A/G	33 (42.3)	39 (51.3)	1.78 (0.88-3.57)	
		A/A	11 (14.1)	14 (18.4)	Dominant (A/A+A/G vs. G/G)	
ATM	rs609429	G/G	31 (39.7)	20 (26.7)	0.10	0.40
		G/C	32 (41.0)	37 (49.3)	1.83 (0.89-3.77)	
		C/C	15 (19.2)	18 (24.0)	Dominant (G/C+C/C vs. G/G)	
ATM	rs609261	T/T	34 (43.6)	23 (29.5)	0.08	0.40
		T/C	33 (42.3)	41 (52.6)	1.87 (0.93-3.75)	
		C/C	11 (14.1)	14 (17.9)	Dominant (C/C+T/C vs. T/T)	
ATM	rs228592	C/C	31 (39.7)	20 (25.6)	0.06	0.40
		C/A	33 (42.3)	40 (51.3)	1.96 (0.96-4.02)	
		A/A	14 (17.9)	18 (23.1)	Dominant (C/A+A/A vs. C/C)	
ATM	rs373759	C/C	41 (52.6)	31 (39.7)	0.13	0.44
		C/T	27 (34.6)	36 (46.1)	1.67 (0.86-3.24)	
		T/T	10 (12.8)	11 (14.1)	Dominant (C/T+T/T vs. C/C)	
ATM	rs611646	T/T	36 (46.1)	24 (30.8)	**0.042**	0.40
		T/A	29 (37.2)	41 (52.6)	**2.04 (1.02-4.08)**	
		A/A	13 (16.7)	13 (16.7)	Dominant (T/A+A/A vs. T/T)	
OGG1	rs293795	A/A	66 (84.6)	70 (89.7)	0.52	0.69
		A/G	12 (15.4)	8 (10.3)	0.72 (0.27-1.94)	
		G/G	0 (0)	0 (0)	A/G vs. A/A	
OGG1	rs1052133	G/G	30 (38.5)	32 (41.0)	0.92	0.94

		G/C	34 (43.6)	31 (39.7)	1.04 (0.45-2.39)	
		C/C	14 (17.9)	15 (19.2)	Recessive (C/C vs. G/C+G/G)	
OGG1	rs2072668	G/G	30 (38.5)	32 (41)	0.92	0.94
		G/C	34 (43.6)	31 (39.7)	1.04 (0.45-2.39)	
		C/C	14 (17.9)	15 (19.2)	Recessive (C/C vs. G/C+G/G)	
MLH1	rs3172297	T/T	67 (85.9)	71 (92.2)	0.16	0.47
		T/C	11 (14.1)	6 (7.8)	0.47 (0.16-1.39)	
		C/C	0 (0)	0 (0)	T/C vs. T/T	
MLH1	rs1800734	A/A	28 (35.9)	24 (30.8)	0.42	0.65
		G/A	40 (51.3)	39 (50.0)	1.46 (0.59-3.63)	
		G/G	10 (12.8)	15 (19.2)	Recessive (G/G vs. G/A +A/A)	
IL-6	rs2069852	A/A	40 (51.3)	39 (50.0)	0.39	0.64
		G/A	29 (37.2)	33 (42.3)	0.62 (0.20-1.89)	
		G/G	9 (11.5)	6 (7.7)	Recessive (G/G vs. G/A+A/A)	
IL-6	rs2069840	C/C	71 (91.0)	68 (87.2)	0.29	0.58
		C/G	7 (9.0)	10 (12.8)	1.78 (0.61-5.22)	
		G/G	0 (0)	0 (0)	C/G vs. C/C	
IL-6	rs2066992	T/T	47 (60.3)	47 (60.3)	0.45	0.65
		G/T	25 (32.0)	27 (34.6)	0.60 (0.16-2.31)	
		G/G	6 (7.7)	4 (5.1)	Recessive (G/G vs. G/T+T/T)	
IL-10	rs3024490	A/A	43 (55.1)	37 (47.4)	0.32	0.58
		A/C	33 (42.3)	32 (41.0)	1.39 (0.72-2.66)	
		C/C	2 (2.6)	9 (11.5)	Dominant (C/C+A/C vs. A/A)	
IL-10	rs1800871	A/A	43 (55.1)	38 (49.4)	0.46	0.65
		A/G	33 (42.3)	30 (39.0)	1.28 (0.67-2.44)	
		G/G	2 (2.6)	9 (11.7)	Dominant (G/G+A/G vs. A/A)	
IL-10	rs1878672	G/G	71 (91.0)	71 (91.0)	0.84	0.90
		C/G	7 (9.0)	7 (9.0)	0.89 (0.28-2.79)	
		C/C	0 (0)	0 (0)	C/G vs. G/G	

GPC5	rs2352028	C/C	43 (55.8)	47 (60.3)	0.59	0.71
		C/T	29 (37.7)	27 (34.6)	0.84 (0.43-1.61)	
		T/T	5 (6.5)	4 (5.1)	Dominant (C/T+T/T vs. C/C)	
TP63	rs10937405	C/C	47 (60.3)	42 (53.9)	0.39	0.64
		C/T	28 (35.9)	33 (42.3)	1.33 (0.69-2.58)	
		T/T	3 (3.8)	3 (3.8)	Dominant (C/C+C/T vs. T/T)	
CY1A1	rs4646903	A/A	22 (28.2)	22 (28.2)	0.81	0.89
		A/G	41 (52.6)	43 (55.1)	0.90 (0.38-2.12)	
		G/G	15 (19.2)	13 (16.7)	Recessive (G/G vs. A/G+A/A)	
CY1A1	rs4646422	C/C	64 (82.0)	65 (83.3)	0.054	0.40
		C/T	12 (15.4)	13 (16.7)	NA	
		T/T	2 (2.6)	0 (0)	Recessive (T/T vs. C/T+C/C)	
CY1A1	rs4646421	G/G	21 (26.9)	22 (30.1)	0.59	0.71
		A/G	42 (53.9)	39 (53.4)	0.82 (0.38-1.73)	
		A/A	15 (19.2)	12 (16.4)	Dominant (A/G+A/A vs. G/G)	
CLPTMIL	rs401681	C/C	34 (43.6)	40 (51.3)	0.45	0.65
		T/C	38 (48.7)	30 (38.5)	0.78 (0.40-1.50)	
		T/T	6 (7.7)	8 (10.3)	Dominant (T/C+T/T vs. C/C)	
CLPTMIL	rs402710	C/C	29 (42.0)	31 (39.7)	0.56	0.71
		T/C	34 (49.3)	39 (50.0)	1.23 (0.61-2.49)	
		T/T	6 (8.7)	8 (10.3)	Dominant (T/C+T/T vs. C/C)	
C3of21	rs2131877	A/A	28 (35.9)	26 (33.3)	0.33	0.58
		A/G	35 (44.9)	42 (53.9)	0.63 (0.25-1.60)	
		G/G	15 (19.2)	10 (12.8)	Recessive (G/G vs. A/A+A/G)	

[a] The P was adjusted by age, gender, family history of lung cancer, family history of other cancer, environmental exposure to smoke at home, environmental exposure to smoke at workplace.

[b] P-BH was adjusted by FDR-BH.

[c] The genetic model included recessive, dominant, additive.

[d] XRCC1 rs1799778; XPC rs1106087, rs3731055, rs2279017; ATM rs664982 were excluded because the call rate<90%; XRCC1 rs3213328; MDM2 rs2279744 were excluded because controls departed from HWE.

Table 3. Validation of two identified SNPs in 84 EGFR mutant lung adenocarcinoma cases compared with 103 CHB and 105 CHS public controls in 1000 genome database.

SNPs	Genotype	84 EGFR Mutant Case N (%)	103 CHB Control N (%)	105 CHS Controls N (%)	P (vs. CHB) OR(95%CI) Genetic model	P' (vs. CHS) OR(95%CI) Genetic model
	G/G	27 (32.1)	35 (34.0)	20 (19.1)	0.99	0.096
ERCC2 rs238406	G/T	35 (41.7)	43 (41.8)	58 (55.2)		
	T/T	22 (26.2)	25 (24.3)	27 (25.7)	Additive (G/T vs. G/G)	Additive (G/T vs. G/G)
	C/C	76 (90.5)	81 (78.6)	92 (87.6)	**0.02**	N.A
IL-6 rs2069840	C/G	8 (9.5)	20 (19.4)	13 (12.4)	**2.76 (1.13-6.72)**	
	G/G	0 (0)	2 (1.9)	0 (0)	Dominant (C/G+G/G vs. C/C)	Dominant (C/G+G/G vs. C/C)

CHB, Chinese Han, Beijing; CHS, Chinese Han, Southern.