UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITÀ

DOTTORATO IN
MEDICINA DEL LAVORO E IGIENE INDUSTRIALE
Direttore: Chiar.mo Prof. Giovanni COSTA

Interaction between genetic and occupational factors in lung cancer etiology. A population-based case-control study.

Tesi di Dottorato di:

Dott. ssa Sara De Matteis

Matricola R08742

Relatore: Chiar.mo Prof. PA. Bertazzi

Correlatore: Dott.ssa Maria Teresa Landi

Anno Accademico 2011-2012

INTRODUCTION

Lung cancer epidemiology

Lung cancer is the leading cause of death due to cancer worldwide: in 2005 World Health Organization (WHO) estimated 1.5 million new lung cancer cases (1.1 million among men and 440,000 among women) (about 12% of total cancer diagnoses) and about 975,000 men and 376,000 women projected to die from it worldwide (1).

Also in European Country lung cancer is the most common cause of cancer death with 334,800 (19.7% of the total), but not the most frequently cancer diagnosed, following breast and colorectal cancer, with 386,300 new cases (12.1% of the total) in the 2006 (2). Specifically, in Italy lung cancer is the first cause of death among men and the second among women accounting for 25,639 and 6,495 deaths respectively in 2002, with a corresponding mean incidence of 111.5 and 27.9 new cases per 100,000 every year (3).

Cigarette smoking is the most important risk factor, accounting for about 80% of lung cancer cases in men and 50% in women worldwide (4).

Occupational risk factors

Among the other environmental risk factors for lung cancer, occupation is the most relevant: since 1972 the International Agency for Research on Cancer (IARC) Monograph Program published almost 100 volumes, reporting evaluations of about 1,000 substances, complex mixtures, and industrial processes (5). In a recent review of these occupational carcinogens, 18 occupations/industries and substances that possibly, probably or definitely (IARC groups 2B, 2A, and 1, respectively), entail an excess risk, with strong evidence for lung cancer, were listed: aluminium production; arsenic and arsenic compounds; asbestos; beryllium; cadmium and cadmium compounds; chromium compounds, hexavalent; coal gasification; coke production; hematite mining, underground, with radon exposure; involuntary (passive) smoking; ionizing radiation; iron and steel founding; selected nickel compounds, including combinations of nickel oxides and sulphides in the nickel refining industry; painters; crystalline silica; soot; talc containing asbestiform fibers (6).

Using the WHO Comparative Risk Assessment (CRA) methodology, the global burden of occupational disease and injury resulting from 8 selected risk factors (beryllium, cadmium, chromium, diesel exhaust, nickel, arsenic, asbestos, silica) in the year 2000 included 850,000 deaths and approximately 24 million years of healthy life lost. Without correction for underestimation, these selected occupational risk factors alone accounted for about 1.5% of all mortality and about 1.6% of all Disability-Adjusted Life Years (DALYs) in the world in the year 2000. The third occupational cause of death was lung cancer (12%) (7).

In Europe, assuming attributable fractions of 7-15% among men and 2-9% among women, 29,300 and 3,200 lung cancer deaths have been estimated respectively (8).

From the epidemiological studies conducted worldwide a great variability in the proportion of lung cancer cases attributable to occupational risk factors emerges, ranging from 0 to 40%, that can be explained with the different proportion of workers exposed to high risk occupations according to time and place specific industrial background (9).

In 1990-1993 the prevalence of working population exposed to occupational carcinogens was still elevated: among the 140 million workers across 15 European Community countries, 32 million resulted exposed and 7 million to the selected 8 lung carcinogens mentioned above (10). In Italy, these estimates were 4 and 1 million respectively and after ten years (2000-2003) only small decreases emerged (11-12).

Genetic risk factors

Even if only 2% of lung cancer cases among males and 30-50% among women have never smoked (13), fewer than 20% of cigarette smokers develop lung cancer (14). Global statistics estimate that 15% of lung cancers in men and 53% in women are not attributable to smoking, overall accounting for 25% of all lung cancer cases worldwide (4). Familial aggregation of lung cancer is reported (15) and a recent meta-analysis estimated a 1.5-fold elevated risk among never smoking probands with affected first degree relatives (16), suggesting that inherited genetic factors may also be important risk determinants.

The research of individual genetic susceptibility for lung cancer has been supported by development of rather simple and rapid new techniques of molecular biology (i.e., polymerase chain reaction (PCR)-based assays) for DNA-sequencing that has enabled precise identification of an individual's genotype.

Beyond the research on uncommon “high-penetrance” genetic mutations, able by themselves (in absence of other factors) to increase lung cancer susceptibility (e.g. KRAS, EGFR, Tp53), highly relevant are the investigations on common (frequency >1% among population) **“low-penetrance”** (increasing risk only in presence of other factors) polymorphisms. In fact, because of the broad occurrence of low-penetrance polymorphisms among general population, the potential impact on public health is pivotal for the possibility to prevent not only the cases attributable to them, but also to interaction with environmental factors (17).

Six biologically plausible patterns of a simple gene-environment interaction model on the relative risk of disease have been proposed (18):

1. **Type 1**: the increased risk of disease is only observed when both genetic and environmental factors co-participate in the same pathogenic mechanism (neither the genotype alone nor the exposure alone causes excess risk);
2. **Type 2**: the environmental exposure is associated with increased disease risk, whereas genotype alone is not;
3. **Type 3**: the genotype is associated with increased disease risk, whereas environmental exposure alone is not;
4. **Type 4**: both the genotype and the environmental exposure are each associated with excess risk of disease with a possible synergistic effect in case of co-exposure;

5-6. **Types 5 and 6**: occur when there is a reversal of the genotype's effect, depending on the presence or absence of environment factors: the genotype is protective in the absence of environmental factors, but is deleterious in the presence of the environmental factors.

A similar, more-complex model, also considering the number of genetic loci involved and of environmental exposure factors, has been proposed (19).

In occupational epidemiology the **Type 2 pattern** is the most studied since **most occupational carcinogens may increase their toxicity in presence of specific metabolic polymorphisms**, but the same **genetic variants are not able by themselves to increase the risk** (20).

Several genes, potentially involved in different carcinogenesis phases, have been intensively studied as suitable “candidates genes” for lung cancer susceptibility, in particular those that would influence lung cancer risk as a result of gene-environment interaction. Genotyping analyses on lung cancer have been carried out on xenobiotic metabolizing enzymes with known genetic polymorphisms, involved in the metabolism of environmental or tobacco carcinogens, and on DNA repair enzymes, involved in repair of DNA damaged by endogenous and exogenous mutagens (21).

Variations in an individual's **metabolic** phenotype, have been detected in a variety of enzymes involved in activation (phase 1: oxidation/reduction/hydrolysis) and detoxification (phase 2: conjugation) of chemical carcinogens. This phenotypic metabolic variation is related to genetic polymorphisms (i.e., metabolic polymorphism). A growing number of genes encoding carcinogen-metabolizing enzymes have been identified and cloned. Consequently, there is increasing knowledge of the allelic variants or genetic defects that give rise to the observed variation (20).

Specifically for lung cancer the most widely studied polymorphic loci are those coding for **phase 1 and 2 enzymes**, involved respectively in the activation of polycyclic aromatic hydrocarbons (PAHs), N-nitrosamines, and aromatic amines and detoxification of epoxides and aromatic amines derived from tobacco smoke. Between them, the most frequently studied enzymes include CYP1A1, microsomal epoxide hydrolase 1 (mEH/EPHX1), myeloperoxidase (MPO), manganese superoxide dismutase (SOD2), NAD(P)H quinone oxidoreductase 1 (NQO1) and the glutathione S-transferases (GST) family, in particular GSTM1 and GSTT1) (22-35).

Nevertheless, the available published data generally offer inconsistent results, likely due to heterogeneity of study populations, failure to consider effect modifiers such as environmental exposures (gene-environment interaction), poor characterization of the exposure, lack of statistical power causing false negatives, and multiple testing creating false-positive results, as well as publication bias (27, 36).

An additional shortcoming of previous studies is that few have focused on detecting the **genetic metabolic polymorphisms** able to increase individual **susceptibility for lung cancer** associated with exposure to **occupational carcinogens,** and they have produced **inconsistent results** (37-42).

STUDY AIM

The aim of this study is to investigate the **interaction** between exposures to **selected known/suspected occupational carcinogens** and **phase II metabolic gene polymorphisms** associated with **lung cancer risk.**

There are several **specific goals**:

1- To improve **understanding** of the mechanisms of action of known or suspected occupational carcinogens in the lung cancer carcinogenesis pathway, for theoretic-scientific purpose;

2- To evaluate the **global impact** of these factors and their interaction on public health, calculated as population attributable fraction (PAF), that estimates the number of cases avoidable every year by eliminating the risk factor in the population exposed;

3- To enable identification of **susceptible subgroups** of the population at higher risk, even at current low exposure levels.

To achieve these aims**,** I have conducted a **candidate gene association study** with **a systematic** and **integrated approach.** To take into account the underlying biological complexity, I adopted a **multi-level approach** that featured analyses at the single nucleotide polymorphism (SNP), gene, haplotype and pathway levels. In addition, I evaluated in gene expression data the correlation between the genetic variants found associated with occupational carcinogens and the genetic functional variants at lung tissue level.

The most important potential impacts of this research would be a re-evaluation of the exposure threshold values that are currently in force, public health campaigns, screening interventions focused on susceptible subjects for primary and secondary prevention (e.g., early cancer detection among exposed workers during health surveillance), with obvious issues also on ethical ground (43-45), and recognition and compensation of occupational cancer cases.

EAGLE (**E**nvironment **A**nd **G**enetics in **L**ung cancer **E**tiology) study, born from the collaboration between the National Cancer Institute (NCI), Bethesda (USA) and the EPOCA research centre of the University of Milan (Italy), is a large population-based case-control study recently conducted in Lombardy region that gives a unique opportunity to achieve these aims: it was designed with the goal of investigating the genetic and environmental determinants of lung cancer, with particular attention to cigarette smoking, using an integrative approach that allows combined analysis of genetic, environmental, clinical, and behavioural data. Moreover, it enrolled a very high number of subjects, also among population controls, and collected detailed information about several important lung cancer determinants and a relevant number of biological samples, obtaining accurate data on exposure and genotype. Besides, given the homogeneous genetic background of the study base (only subjects born in Italy, with Italian citizenship and residence in Lombardy region) there's a minimal possibility of confounding by different genetic backgrounds within ethnic groups (i.e., population stratification).

MATERIAL AND METHODS

EAGLE Study: population and data collection

A detailed description of the EAGLE study has been previously published (46). Briefly, the study includes 2,100 incident lung cancer cases and 2,120 population controls enrolled in the period April 2002 to June 2005 in 216 municipalities in the Lombardy region (Northern Italy). Cases were subjects with primary cancer of trachea, bronchus, and lung, first diagnosed between April 2002 and February 2005, and admitted to 13 hospitals with catchment of greater than 80% of the lung cancer cases in the study area. Controls were randomly sampled from the Regional Health Services Database, frequency-matched to cases by area of residence (5 classes), gender, and age (5-year categories), and contacted through the family physician. All enrolled subjects were Caucasian. Subjects were 35–79 years of age at diagnosis (cases) or at sampling/enrolment for interview (controls). The study participation rates were 86.6% among cases and 72.4% among controls. After signing an Institutional Review Board-approved informed consent form, subjects underwent a computer-assisted personal interview (CAPI) and filled-in a self-administered questionnaire. Available data includes demographical characteristics, detailed smoking history, family history of lung cancer and other cancers, previous lung diseases, medications, diet, alcohol, attempts at quitting smoking, anxiety, depression, personality scores, occupations, reproductive and residential history.

Particular attention was given to the collection of data on tobacco exposure including active smoking (age at initiation/cessation, number of cigarettes per day in different periods) and passive smoking (during childhood, at work, and at home during adulthood).

Clinical data (stage, grade, histology, imaging and pathology reports, spirometry, and routine laboratory tests) were recorded. All study subjects donated a blood sample (or, rarely, a buccal rinse sample), which was processed to obtain cryopreserved lymphocytes, red blood cells, granulocytes, DNA, RNA, whole blood, buffy coat, serum, plasma, and blood cards. Lung tissue paraffin blocks and slides were collected from the cases that underwent surgery, biopsy or cytological examination of the lung tumor. Multiple fresh tumor and “non-involved” lung tissue samples, frozen in liquid nitrogen within 20 minutes of excision, were also collected from over 500 surgical cases.

Epidemiological and biospecimen information has been collected respectively for 98.4% and 97.3% of cases and 99.8% and 99.9% of controls, then anonymized and stored in a secure relational database. Quality control procedures were implemented to ensure data completeness and accuracy. Several genetic and epidemiological studies are ongoing.

Occupational exposure assessment

For jobs held for at least six months, detailed information on lifetime work history (industry, job title, year of start and stop) was collected for the 1,943 cases and 2,116 controls that underwent CAPI. Jobs (industry and job title) were then coded, blindly with respect to case-control status, by occupational physicians with training and experience in epidemiology and industrial hygiene, by using the International Standard Industrial Classification of All Economic Activities (ISIC), Revised Edition 2 (47), and the International Standard Classification for Occupations (ISCO), 1968 (48).

In absence of a gold standard different approaches for occupational exposure assessment can be used, each with advantages and limits (49-51). In the present work I have applied a general Job-Exposure Matrix (JEM) to estimate the individual exposure to selected occupational carcinogens. The JEM approach, besides being very cheap and easy to apply, allows converting each job code into the specific exposures entailed by it, gathering workers with common exposure irrespective of their occupational titles, so increasing categorization sensitivity. The important advantage is the possibility of evaluating the causal role of single occupational carcinogens known or suspected to be associated to lung cancer. The main limit is the potential non-differential misclassification due to the heterogeneity of the industries/occupations combinations grouped in the same exposure categories, with consequent underestimation of the risk effect (52-54).

JEM

The JEM used in this study is the ‘**DOM-JEM’**, recently developed within the SYNERGY project, an international pooled analysis of lung cancer case–control studies coordinated by the International Agency for Research on Cancer (IARC), the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-University Bochum (IPA) and the Institute for Risk Assessment Sciences at Utrecht University (IRAS) (http://synergy.iarc.fr). This semi-quantitative JEM was created *a priori* (i.e. independently from any study population) to be applied in community-based studies. Experts' rating was based on intensity and probability of exposure (55). The JEM translates all job titles (five-digit ISCO codes) into exposure to selected agents, ranked as 0, 1 and 2 for no, low and high exposure, respectively.

The six known/suspected occupational lung carcinogens included in the 'DOM-JEM' were **asbestos, crystalline silica, polycyclic aromatic hydrocarbons (PAH), diesel motor exhausts (DME), chromium compounds (Cr)** and **nickel compounds (Ni).** These agents had been previously selected for the SYNERGY project, according to the following criteria: (i) IARC evaluation: known (Group 1) or suspected (Group 2A/2B) lung carcinogens; (ii) relevance for recognition of occupational diseases associated with these agents (recognized number of cases per year by Workers Health Insurance); (iii) prevalence of exposure and probability of simultaneous exposures to two or more agents over the course of an individual job history in the general population; and (iv) available information for quantitative exposure assessment. I merged the five-digit ISCO codes for jobs held by each subject with the JEM to estimate the individual exposures.

Genetic analysis

Candidate gene approach

For the genetic analysis I used a **candidate gene approach** to test directly the interaction of selected genetic polymorphisms and the occupational carcinogens included in the JEM in association with lung cancer risk. The main advantage of this method is that it is the most powerful in a population-based case-control study to detect the small effect of low-penetrance genes in association with complex disease traits like lung cancer. Another important advantage is that it is relatively cheap and quick. The major drawback is that the incomplete knowledge of the underlying biological mechanism limits the number of genes that can be tested to the ones for which at least some functional information is available (56).

Candidate genes selection

I conducted a comprehensive **review** of the literature available on this topic (22-35, 37-42, 57-59) and I selected the candidate **phase II metabolic genes** that have been reported in association with:

1) Lung cancer susceptibility
2) The metabolism of the 6 occupational carcinogens included in the JEM
3) Lung cancer risk and exposure to the 6 occupational carcinogens included in the JEM

This is the final list of **23 candidate genes** that were evaluated in this study:

ABCG2, ALDH2, CAT, COMT, GSTA1, GSTA2 ,GSTA3, GSTA4, GSTCD, GSTM2, GSTM3, GSTM4, GSTM5, GSTP1, GSTT2, GSTZ1, MDR1, MPO, NAT1, NAT2, NQO1, SOD2, UGT1A7.

The GSTT1 and GSTM1 genes were not included in this analysis because previously evaluated in the EAGLE study in a work recently published (35).

SNP selection

The SNP selection for the EAGLE study had been previously described (31). SNP assays were selected from those available at the Core Genotyping Facility (CGF) of the Division of Cancer Epidemiology and Genetics (National Cancer Institute), using NCI assessment of linkage disequilibrium (LD) (i.e., the non-random association of alleles at two or more loci) between the SNPs from the International Haplotype Mapping Project (HapMap) database which contains the LD patterns of European, African and Asian populations and previous evidence from the literature.

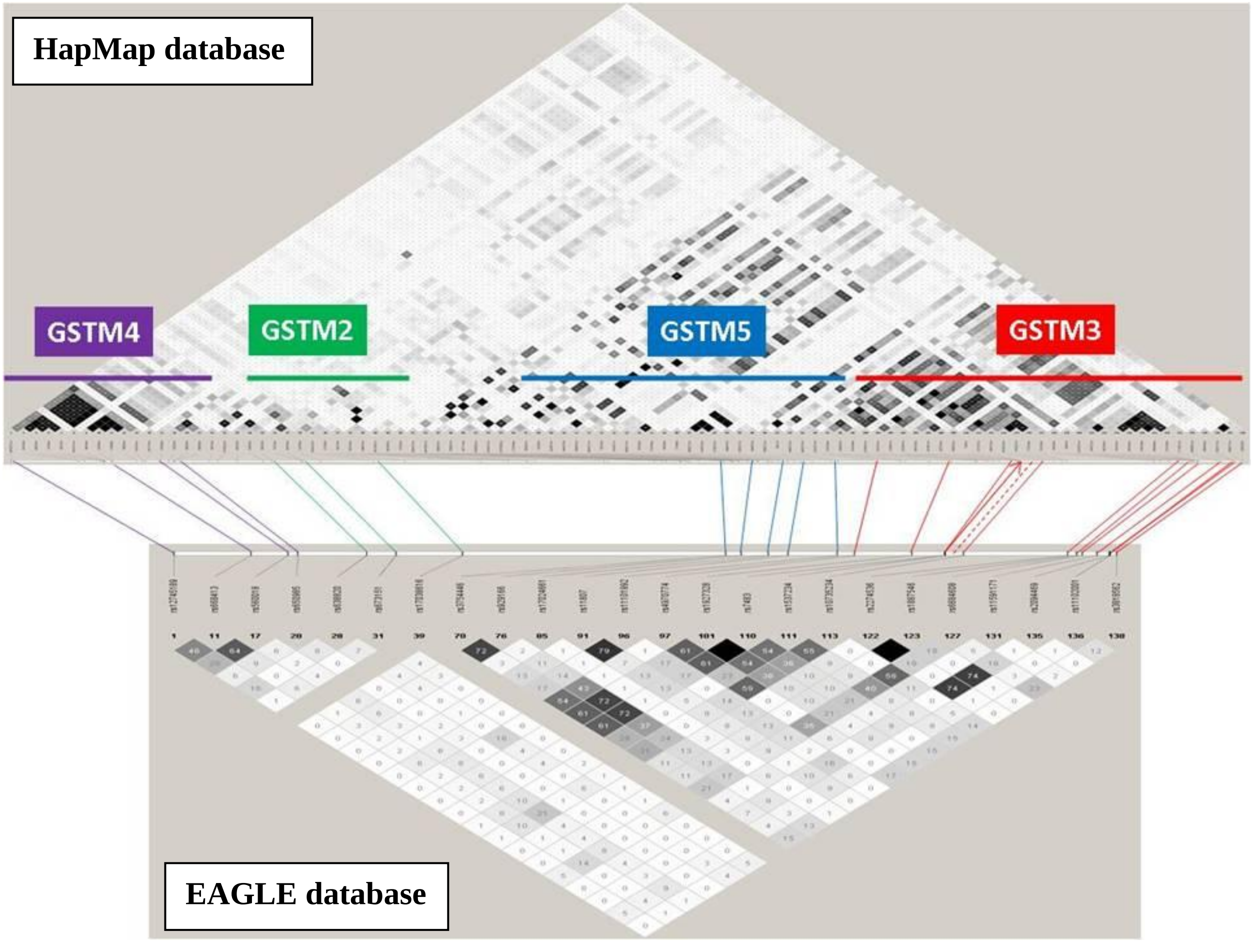
I selected **298 tagging SNPs**.

Gene coverage

For all the **23 candidate genes,** represented in the data by two or more SNPs, I evaluated the genetic coverage of the **298 selected tagging SNPs** using Haploview software to estimate and visualize the **pairwise LD**, using as reference the present version of the HapMap database.

An example of the **good coverage** of the selected tagging SNPs for the **GSTM family genes** is shown below (**Graph 1**)**.**

Graph 1. Gene coverage of selected tagging SNPs for GSTM family genes in the EAGLE study compared to Hap Map reference dataset.



SNP genotyping

Genotyping of the selected 298 SNPs was performed on all the 4,050 EAGLE subjects with sufficient DNA samples, followed by quality-control procedures, and conducted at CGF of NCI using two types of assays: customized TaqMan® probes described at the NCI SNP500Cancer website (http://snp500cancer.nci.nih.gov) and standard Illumina HumanHap550v3_B BeadChips (Illumina, San Diego, CA, USA). The aim was to combine the specificity of the first assay with the sensitivity of the second in order to increase the genetic coverage. In case of duplicates between assays, the SNPs from TaqMan were retained for the analyses, because they were more specific for the selected genes.

Gene expression data

Data on microarray gene expression from peripheral blood lymphocytes were obtained using the Affymetrix GeneChip® HG-U133A v2.0, already described in detail (35). Briefly, the samples were processed and normalized with the Robust Multichip Average (RMA) method. All 22,277 probe sets based on RMA summary measures were used in the analyses. For the present study I used data from paired tumor (n = 51) and non-involved (n = 41) lung tissue samples from lung cancer cases and from peripheral whole blood of cases (n = 71) and controls (n = 76) with available data on occupational exposure.

Statistical analysis

Occupational exposure

For each carcinogen, I evaluated a dichotomous exposure indicator (never/any), and an ordinal variable for intensity of exposure (never/low/high). Further, I analysed duration and cumulative exposure as the sum of the job-specific (intensity score × duration) products (with scores set to 1 and 4 for low and high exposure, respectively). Latency was defined as time at lung cancer diagnosis or study enrolment since first exposure. The analyses were conducted using both categorical and continuous variables. For duration and cumulative exposure I defined the categories according to the quartiles of the exposure distribution among controls for each carcinogen. For latency I used predefined categories of exposure (never, 20-29, 30-39, 40-49, 50-59 and ≥60 years) to explore their impact on a broader range of years since first exposure. When analysing those variables as continuous, I used the ln (1 + x) transformation to normalize their distribution. I evaluated co-exposure to the JEM carcinogens using Spearman's rank correlation coefficient (ρ_s).

For each carcinogen exposure I calculated odds ratios (ORs), 95% confidence intervals (95% CIs) and tests for trend, using unconditional logistic regression, separately for males and females, taking subjects never exposed to the carcinogen as reference. All regression models included the following covariates: residential area (five categories); age (five-year categories); cigarette smoking (ever/never); pack-years (continuous, mean-centred: linear, quadratic, and cubic terms); time since quitting (0 for never/current smokers, 0.5, 1, 2, 5, 10, 20, ≥30 years); smoking (ever/never) of other types of tobacco (pipe, cigars, cigarillos); and, for each agent, co-exposure to the other carcinogens included in the JEM. I also adjusted for number of jobs held (1, 2, 3, 4, ≥5), since this variable was negatively associated with lung cancer among non-exposed subjects (P_{trend}=0.014) and positively associated with exposure to carcinogens among controls (P <0.0001 from chi-squared test). I repeated selected analyses after adjusting for education (none, elementary, middle, and high school/higher degree) as a surrogate of socioeconomic status.

For the exposures showing an increased OR, I calculated the carcinogen-specific and overall PAF by using the formula $P_{EC} \times (\text{OR} - 1)/\text{OR}$, where OR is the adjusted OR and P_{EC} is the proportion of cases ever exposed to the carcinogen under study (60). The definition of exposure I used when calculating PAF estimates considers subjects unexposed to the carcinogen under study as belonging to the “reference” category and everyone even slightly exposed as belonging to the “exposed” category. Estimates of PAF when using this broad definition of “exposed” are less prone to bias from non-differential misclassification of exposure, the form of misclassification expected with a JEM approach (61). I estimated ORs for the three main histological lung cancer types (adenocarcinoma, squamous cell, and small cell carcinomas) and tested their homogeneity in a multinomial logistic regression model.

I evaluated interactions between each carcinogen (never/any exposure) and cigarette smoking status (never/former/current) on the multiplicative scale, by comparing the likelihood of a logistic regression model containing the main effects of the carcinogen and smoking, with that of a model with also their interaction. As reference, I used subjects never exposed to both smoking and the specific carcinogen under study. In these models I did not adjust for co-exposure to the other JEM carcinogens to avoid too few subjects per strata.

All *P* values were two-sided. Analyses were performed with Stata11 (62). Confidence limits of PAF were calculated with the command aflogit which implemented the formulas proposed by Greenland and Drescher (63).

Genetic main effect

Single SNP analysis

The main effect of the variant genotypes on the risk of lung cancer was estimated by ORs and their 95% CIs using unconditional logistic regression analysis for all subjects and separately by gender. Homozygosity for the more frequent allele among controls was defined as the reference group (AA). I tested for significance using two-sided Wald tests. I evaluated the SNP effect both as continuous variable and as three levels categorical variable to test for linear trends. I adjusted the ORs for the matching variables (age, sex, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.

In all the analyses I evaluated **three models of genetic inheritance**:

1. **Additive model**: the risk conferred by an allele is increased r-fold for heterozygotes and 2r-fold for homozygotes. This model assumes a linear relationship between the number of allele copies and the associated trait, allowing performing test for trends.
2. **Dominant model**: the risk conferred by an allele (dominant allele) is the same for heterozygotes and homozygotes. The comparison groups are wild-type homozygous genotypes vs. allele positivity (combining heterozygotes and homozygotes for the variant).
3. **Recessive model:** the risk conferred by an allele (recessive allele) is present only for homozygotes. The comparison groups are variant homozygous genotypes vs. the rest (combining heterozygotes for the variant and homozygotes for the wild-type allele).

Assuming that a SNP has genotypes of AA (“wild-type” homozygote), AB (heterozygote) and BB (variant homozygote), the genotypes were coded as AA=0, AB=1 and BB=2 in an additive model; AA=0 AB=1 and BB=1 in a dominant model; and AA=0, AB=0 and BB=1 in a recessive model.

Also, I estimated ORs for the three main histological lung cancer types (adenocarcinoma, squamous cell, and small cell carcinomas) and tested their homogeneity in a multinomial logistic regression model as reported above for the occupational exposure analysis.

SNP grouped by genes analysis

I analysed multiple SNPs jointly to test whether the overall lung cancer risk was determined by the combined action of multiple SNPs within the same gene and/or of multiple genes within the same pathway, even if each SNP may have had only a modest effect individually.

For these analyses I tried to increase the statistical power by excluding from the dataset the “redundant” SNPs, i.e., the SNPs in high LD more likely to be transmitted together and so carrying the same genetic information. I estimated the LD between the diallelic SNPs among controls and I excluded the SNPs with r^2 >0.80.

For the SNP grouped analysis I used two models:

1) **SNP grouped "cumulative" analysis**

Under the assumption that the effect on lung cancer of each SNP was cumulative, I implemented the following logistic regression model:

$$Logit\ (LC) \quad = \alpha + \beta \ \times \sum_{k}^{n} (SNP_k) + \gamma \ \times covariates$$

where k=1, …, n represents a collection of SNPs belonging to the same gene or a collection of SNPs belonging to genes in the same pathway (e.g. phase II, n = 23 i.e. all SNPs were grouped together). SNP_k = 0 for the homozygote most common allele, SNP_k= 1 for the heterozygote allele, and SNP_k = 2 for the homozygote minor allele. β is the regression coefficient for the **cumulative number of variants** $\sum_{k}^{n}$ (SNP_k).

I estimated the overall risk of lung cancer (LC in the formula above) associated with each selected group of n SNPs by computing OR = exp (β). Note that in this model I do not assume nor infer a risk direction for each minor allele. This approach is powerful if minor alleles for all SNPs have effects in the same direction, but there may be loss of power if minor alleles for some SNPs affect lung cancer risk in opposite directions and their contribution to the overall risk cancels with each other.

2) **SNP grouped "score" analysis**

For this analysis, in the same logistic model explained above, I treated the **cumulative sum of the effect** of each SNP (expressed by the regression coefficients β) within each gene as independent variable, as shown below:

$$Logit\ (LC) \quad = \alpha + \beta \ \times \sum_{k}^{n} (\beta_k) + \gamma \ \times covariates$$

Genetic-occupational interaction

Applying the same unconditional logistic regression model used for the main effect of genetic variants, I calculated ORs and 95%CIs by exposure to each occupational carcinogen (dichotomous variable: ever/never exposure), using as reference category the subjects with the "wild type" genetic variant who have never been exposed to the carcinogen under evaluation.

Then I tested the interaction between genetic variants and exposure to occupational carcinogens on the multiplicative scale by using a 2-df likelihood ratio test (LRT) comparing the logistic regression model containing only the main effect of genetic and occupational variables and a model containing also their interaction effect.

Multiple comparison considerations

Given the high number of hypotheses tested in the single SNP analyses (298 tests corresponding to the 298 SNPs for the single SNP analysis and 23 tests when SNPs were grouped by genes), I took multiple testing into account. I chose the Benjamini-Hochberg (64) procedure to calculate the False Discovery Rate (FDR) in preference to the more conservative Bonferroni correction (i.e. testing each of the individual tests at a significance level of α/n, where α is the statistical significance threshold, and n is the number of performed tests).

In fact, my approach to multiple testing was "informed" by the selection strategy for the phase II genes selected. As previously reported in the Methods section, each of the genes included has substantial mechanistic and at least some population data which support an association with lung cancer and/or occupational carcinogens. I recognize that considering this to be *a priori* knowledge for each SNP may be open to debate, because of the heterogeneity of results in the literature and because most results actually refer to genes and not to specific SNPs, however it was not my aim to perform an explorative and totally "agnostic" analysis.

I considered significant those results with a FDR-corrected-p-values ≤0.05. In addition, I referred to results with p-values between 0.01 and 0.05 as nominally significant, and considered them as notable when consistent across different analyses.

Pathway analysis

To take into account the complex interaction between genes involved in same biological function I performed a pathway analysis. I evaluated pathways that had been defined in externally curate databases (e.g., HuGE, KEGG, BioCarta, PID, etc.) and that have been previously evaluated in association with the outcome of interest (65-68). I used an approach combining gene-level P-values across the candidate genes included in the selected biological pathway through an adaptive rank-truncated product (ARTP) method that uses a permutation algorithm for the evaluation of its significant level (69).

I evaluated **6 pathways:**

1. **GSTM:** GSTM2, GSTM3, GSTM4, GSTM5.
2. **GSTA:** GSTA1, GSTA2, GSTA3, GSTA4
3. **NAT:** NAT1, NAT2
4. **ANTIOXIDANT:** SOD2, CAT
5. **GST:** GSTM, GSTCD, GSTA, GSTP1, GSTZ1, GSTT2
6. **PHASE II METABOLISM**: ALL 23 CANDIDATE GENES

These pathways were tested for association with lung cancer among never and ever exposed to each occupational carcinogen.

Haplotype analysis

To take into consideration that biologically on the same chromosome at each genetic locus there are two haplotypes (i.e., the combination of alleles inherited, one maternally and the other paternally) I performed an haplotype analysis using the haplo.stats R-package that infers haplotype frequencies by assuming that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic markers). The genetic markers are assumed to be co-dominant (i.e., one-to-one correspondence between their genotypes and their phenotypes). Because there may be more than one pair of haplotypes that are consistent with the observed marker phenotypes, posterior probabilities of pairs of haplotypes for each subject were also computed using a "progressive insertion" algorithm which progressively inserts batches of loci into haplotypes of growing lengths, runs the expectation–maximization (EM) steps, trims off pairs of haplotypes per subject when the posterior probability of the pair is below a specified threshold, and then continues these insertion, EM, and trimming steps until all loci are inserted into the haplotype. Only the haplotypes with a frequency above 0.02 were included in the analysis.

I tested in the same regression model used in the previous analysis the haplotype-carcinogen interaction term for each gene evaluated using the most frequent haplotype among controls as reference.

Gene expression analysis

Limited to the genes found significantly associated with exposure to occupational carcinogens for lung cancer risk to better understand the underlying biological mechanism, I estimated the effect of each SNP from a given gene on the expression of the same gene in the four types of tissue mentioned above. I evaluated the correlation between the number of genetic variations and mRNA expression using linear models (i.e., log2 expression = α + ß x genetic variant) adjusted for the same covariates included in the other analyses and computing fold changes (FC = $2^{ß}$) of expression between individuals with different genetic variants among never and ever exposed to the occupational carcinogen under evaluation.

Statistical software

All statistical analyses of genetic data were performed using the **Rproject** (version 2.10) statistical package (http://www.r-project.org/index.html).

RESULTS

Study base characteristics

The frequency distributions for the main covariates among the 4,016 subjects included in the EAGLE study are shown in the **Table 1**. Of the 2,100 cases and 2,120 controls enrolled in our study, 1,943 (92.5%) and 2,116 (99.8%) were interviewed, respectively. Two-thirds of the subjects came from the Milan area. Among men, controls had higher education and held more jobs than cases. About 14-

15% of cases and 6-7% of controls had previously or newly-diagnosed primary cancer(s) other than lung cancer. Among cases, one-fourth of women were never smokers, versus only 2% of men. In both genders, current smokers were around 50% among cases and less than 30% among controls. Almost half of men (cases or controls) were former (quit > six months ago) smokers, compared to less than 30% among women. The majority of lung cancers were adenocarcinomas (>50% in women).

Table 1. Selected characteristics of lung cancer cases and controls with interview data available, the EAGLE study, Lombardy, Italy, 2002–2005.

	Women				**Men**			
	Cases		**Controls**		**Cases**		**Controls**	
	N	**%**	***N***	**%**	***N***	**%**	***N***	**%**
Total participants enrolled	448		500		1652		1620	
Interviewed	406	100.0	499	100.0	1537	100.0	1617	100.0
Area of residence								
Milan	288	70.9	349	69.9	987	64.2	1089	67.3
Monza	24	5.9	23	4.6	109	7.1	94	5.8
Brescia	47	11.6	53	10.6	203	13.2	194	12.0
Pavia	21	5.2	37	7.4	107	7.0	92	5.7
Varese	26	6.4	37	7.4	131	8.5	148	9.2
		P = 0.55				*P* = 0.17		
Age (years)								
Mean (SD)	64.8	(10.1)	64.1	(10.1)	66.8	(7.9)	65.8	(8.1)
		P = 0.32				*P* < 0.001		
Education level								
None	21	5.2	24	4.8	91	5.9	66	4.1
Elementary	128	31.5	143	28.7	625	40.7	431	26.7
Middle	134	33.0	158	31.7	424	27.6	455	28.1
High	104	25.6	135	27.1	314	20.4	441	27.3
University	19	4.7	39	7.8	83	5.4	224	13.9
		P = 0.35				*P* < 0.001		
Number of jobs								
1	166	40.9	168	33.7	375	24.4	370	22.9
2	96	23.7	158	31.7	404	26.3	356	22.0
3	77	19.0	82	16.4	305	19.8	356	22.0
4	30	7.4	49	9.8	194	12.6	226	14.0
5+	37	9.1	42	8.4	259	16.9	309	19.1
		P = 0.03				*P* = 0.02		
Cigarette smoking								
Never	103	25.4	282	56.5	29	1.9	397	24.6
Former (quit >6 months ago)	116	28.6	110	22.0	723	47.0	799	49.4
Current	187	46.1	107	21.4	785	51.1	420	26.0
Unknown	0	0.0	0	0.0	0	0.0	1	0.1
		P < 0.001				*P* < 0.001		
Cigarette pack-years								
Mean (SD)	24.3	(23.1)	7.2	(13.5)	50.9	(28.7)	22.1	(23.2)
		P < 0.001				*P* < 0.001		
Other cancer(s)[c]								
No	336	82.8	448	89.8	1306	85.0	1473	91.1
Yes	70	17.2	51	10.2	231	15.0	144	8.9
		P = 0.002				*P* < 0.001		
Lung cancer morphology								
Adenocarcinoma	220	54.2			582	37.9		
Squamous cell carcinoma	45	11.1			459	29.9		
Large cell carcinoma	28	6.9			61	4.0		
Non-small cell carcinoma NOS	34	8.4			142	9.2		
Small cell carcinoma	38	9.4			157	10.2		
Others	26	6.4			65	4.2		
Not available	15	3.7			71	4.6		
				P < 0.001				

Abbreviations: EAGLE, Environment And Genetics in Lung cancer Etiology; NOS, not otherwise specified; SD, standard deviation. [a]*P* values were derived from the χ^2 test (categorical variables) or Student's t test (continuous variables) between cases and controls. [b]Percentages may not add to 100.0 because of rounding. [c]Primary cancer(s) (previously or newly diagnosed) other than lung cancer.

Occupational exposure

The results of the occupational exposure analysis have been recently published (70) and presented as a talk at the 22nd International Conference on Epidemiology in Occupational Health (EPICOH) in Oxford, UK, September 7th, 2011.

Briefly, men were most commonly exposed to asbestos (41.1% among cases and 32.2% among controls) and DME (38.8% among cases and 38.5% among controls). Intensity levels for the majority of exposed subjects were low.

In the regression model adjusted for area, age, cigarette smoking, other types of tobacco and number of jobs held, we found increased ORs for lung cancer for any and even low exposure to asbestos, silica and Ni–Cr, with positive trends for intensity of exposure. For PAH, only subjects with high exposure had an increased risk. After adjusting also for co-exposure to the other JEM carcinogens, the estimates for associations tended to decrease for all carcinogens, in particular for high exposure to asbestos, low exposure to Ni–Cr and high exposure to PAH. No association was found for DME.

The **PAFs** for any exposure to **asbestos, silica** and **Ni–Cr** were **18.1%, 5.7%,** and **7.0%,** respectively, corresponding to an overall PAF of 22.5% (95% CI: 14.1–30.0) (**Table 2**).

Table 2. Lung cancer risk for exposure to the six job–exposure matrix carcinogens for men in the EAGLE study, Lombardy, Italy, 2002-2005.[a]

	Cases		**Controls**		**OR[b]**	**95%CI**	**OR[c]**	**95%CI**	**PAF[d]**	**95%CI**
	N	**%**	***N***	**%**					**%**	
Asbestos										
Never[e]	905	58.9	1097	67.8	1.00		1.00			
Any	632	41.1	520	32.2	1.73	1.43, 2.09	1.78	1.46, 2.18	18.1	12.6, 23.3
Low	546	35.5	448	27.7	1.68	1.38, 2.04	1.76	1.42, 2.18		
High	86	5.6	72	4.5	2.09	1.39, 3.13	1.51	0.94, 2.44		
					P < 0.001		*P* < 0.001			
Silica										
Never[e]	1166	75.9	1363	84.3	1.00		1.00			
Any	371	24.1	254	15.7	1.38	1.10, 1.72	1.31	1.02, 1.68	5.7	0.4, 10.6
Low	328	21.3	226	14.0	1.37	1.09, 1.73	1.31	1.00, 1.71		
High	43	2.8	28	1.7	1.46	0.81, 2.61	1.41	0.77, 2.55		
					P = 0.006		*P* = 0.02			
Ni-Cr										
Never[e]	1041	67.7	1216	75.2	1.00		1.00			
Any	496	32.3	401	24.8	1.41	1.16, 1.72	1.28	1.00, 1.63	7.0	0.2, 13.3
Low	370	24.1	328	20.3	1.33	1.08, 1.65	1.18	0.90, 1.53		
High	126	8.2	73	4.5	1.77	1.22, 2.56	1.31	0.86, 1.97		
					P < 0.001		*P* = 0.06			
PAH										
Never[e]	1137	74.0	1235	76.4	1.00		1.00			
Any	400	26.0	382	23.6	1.11	0.90, 1.36	0.87	0.68, 1.10		
Low	284	18.5	321	19.9	0.90	0.72, 1.13	0.78	0.61, 1.00		
High	116	7.5	61	3.7	2.46	1.65, 3.67	1.64	0.99, 2.70		
					P = 0.007		*P* = 0.75			
DME										
Never[e]	940	61.2	994	61.5	1.00		1.00			
Any	597	38.8	623	38.5	0.90	0.75 , 1.09	0.82	0.67 , 1.00		
Low	476	31.0	500	30.9	0.89	0.73 , 1.09	0.85	0.69 , 1.05		
High	121	7.8	123	7.6	0.96	0.68 , 1.35	0.70	0.48 , 1.00		
					P = 0.44		*P* = 0.047			

Abbreviations: CI, confidence interval; DME, diesel motor exhausts; EAGLE, Environment And Genetics in Lung cancer Etiology; Ni-Cr, nickel and chromium compounds; OR, odds ratio; PAF, population attributable fraction, PAH, polycyclic aromatic hydrocarbons. [a]*P* values were calculated from test for linear trend for never/low/high exposure. [b]OR calculated with unconditional logistic regression models, adjusted for area, age, smoking, and number of jobs. [c]OR adjusted as [b] and also for co-exposure to the other job–exposure matrix carcinogens. [d]PAF calculated for any exposure to each carcinogen associated to an increased risk using [c] OR and % of cases exposed to each carcinogen. [e]Reference category: never exposed to the specific carcinogen.

Given that **asbestos** was the carcinogen with the **highest impact** in our study in terms of both prevalence of exposure and strength of association with lung cancer risk, I decided to test the interaction between the 23 selected candidate genes and asbestos exposure only in this study. The other five carcinogens included in the JEM will be evaluated in future studies.

Genetic analysis

SNP analysis

Among the 298 SNPs (19 from TaqMan assay and 279 from GWAS chip) tagging 23 phase II metabolic genes potentially involved in asbestos detoxification process, I found 5 duplicates between the assays: rs7483 (GSTM3), rs1001179 (CAT), rs1695 (GSTP1), rs1138272 (GSTP1), and rs4680 (COMT). I chose the SNPs genotyped with the TaqMan assay, as stated in the Methods section, so the final number of SNPs evaluated was reduced to **293**.

All analyses were restricted to the **3,899 subjects** with at least a 90% genotype call rate. All 293 SNPs passed the test for Hardy-Weinberg equilibrium genotype proportions among the 2,041 controls, with a p-value of 0.05 as the threshold.

The frequency of subjects in the EAGLE study with genotype and asbestos exposure data available are shown in the **Table 3**: The four tagging SNPs for the **GSTM4 gene** are shown as an example.

Table 3. Frequency of subjects with genotype and asbestos exposure (ever/never) data available, in the EAGLE study, Lombardy, Italy, 2002-2005. The four tagging SNPs for the GSTM4 gene are reported as an example.

GSTM4 SNP Name	**Genotype**	**Controls**	**Cases**	**Controls Never Exposed**	**Controls Ever Exposed**	**Cases Never Exposed**	**Cases Ever Exposed**
rs12745189	SNP = 0	553	540	403	150	325	181
	SNP = 1	991	913	727	261	554	303
	SNP = 2	434	464	308	125	290	147
	SNP = NA	142	183	108	34	96	47
rs668413	SNP = 0	734	717	522	209	458	209
	SNP = 1	955	889	698	256	536	307
	SNP = 2	290	314	218	72	176	117
	SNP = NA	141	180	108	33	95	45
rs560018	SNP = 0	890	855	633	254	541	257
	SNP = 1	858	824	638	219	492	288
	SNP = 2	217	223	155	62	124	84
	SNP = NA	155	198	120	35	108	49
rs650985	SNP = 0	1798	1780	1311	484	1075	600
	SNP = 1	177	136	123	53	91	33
	SNP = 2	3	3	3	0	3	0
	SNP = NA	142	181	109	33	96	45

Abbreviations: NA = Not Available.

SNP has genotypes coded as 0 for “wild-type” homozygote, 1 for heterozygote, and 2 for variant homozygote.

Given the large amount of test performed in the following tables I have reported the results only for the SNPs found nominally (in italics) or statistically (in bold) associated with asbestos exposure for lung cancer risk. Both raw and FDR corrected p-values are reported.

All subjects: Considering an additive model, the SNPs **rs668413** and **rs560018** tagging the **GSTM4 gene** showed a null effect among never exposed to asbestos, and an increase risk among ever exposed with a positive trend per allele copy (p_{trend} values=0.002 and 0.006, respectively). The LRT p-values for interaction with asbestos exposure ($p_{interaction}$ values =0.004 and 0.015, respectively) did not remained statistically significant after the FDR correction for multiple comparison (**Table 4**).

Table 4. ORs and 95% CIs of lung cancer for SNPs by never/ever asbestos exposure for significant (bold) or nominally significant (italics) SNP-asbestos interactions in the EAGLE study, Lombardy, Italy, 2002-2005. All subjects.

Gene	SNP Name	Comparison	Co Nev Asb	Ca Nev Asb	Nev OR	Nev CI1	Nev CI2	Nev p-value	Co Ever Asb	Ca Ever Asb	Ever OR	Ever CI1	Ever CI2	Ever p-value	LRT p-value	LRT FDR p-value
up GSTM4	rs12745189	1) Trend	403	325	1.04	0.92	1.19	0.523	150	181	0.95	0.78	1.15	0.587	0.419	0.909
up GSTM4	rs12745189	2) AA=0, AB=1	727	554	0.88	0.71	1.10	0.259	261	303	0.86	0.63	1.19	0.368	0.660	0.920
up GSTM4	rs12745189	3) AA=0, BB=1	308	290	1.10	0.85	1.43	0.448	125	147	0.91	0.62	1.33	0.627	NA	NA
up GSTM4	rs12745189	4) AA=0, AB+BB=1	1035	844	0.95	0.77	1.16	0.611	386	450	0.88	0.65	1.19	0.396	0.676	0.997
up GSTM4	rs12745189	5) AA+AB=0, BB=1	308	290	1.20	0.96	1.48	0.105	125	147	1.00	0.72	1.38	0.991	0.362	0.882
up GSTM4	*rs668413*	*1) Trend*	*522*	*458*	*0.96*	*0.84*	*1.10*	*0.568*	*209*	*209*	*1.36*	*1.12*	*1.66*	*0.002*	*0.004*	*0.625*
up GSTM4	rs668413	2) AA=0, AB=1	698	536	0.85	0.7	1.04	0.106	256	307	1.26	0.93	1.70	0.135	0.016	0.598
up GSTM4	rs668413	3) AA=0, BB=1	218	176	1.00	0.75	1.31	0.977	72	117	1.93	1.28	2.93	0.002	NA	NA
up GSTM4	rs668413	4) AA=0, AB+BB=1	916	712	0.88	0.73	1.06	0.190	328	424	1.40	1.05	1.85	0.020	0.008	0.751
up GSTM4	rs668413	5) AA+AB=0, BB=1	218	176	1.09	0.85	1.41	0.504	72	117	1.70	1.16	2.48	0.006	0.056	0.730
GSTM4	*rs560018*	*1) Trend*	*633*	*541*	*0.98*	*0.85*	*1.12*	*0.749*	*254*	*257*	*1.32*	*1.08*	*1.62*	*0.006*	*0.015*	*0.874*
GSTM4	rs560018	2) AA=0, AB=1	638	492	0.89	0.73	1.08	0.233	219	288	1.52	1.13	2.03	0.005	0.010	0.590
GSTM4	rs560018	3) AA=0, BB=1	155	124	1.06	0.78	1.44	0.727	62	84	1.55	1.00	2.42	0.051	NA	NA
GSTM4	rs560018	4) AA=0, AB+BB=1	793	616	0.92	0.77	1.11	0.374	281	372	1.52	1.16	2.01	0.003	0.003	0.751
GSTM4	rs560018	5) AA+AB=0, BB=1	155	124	1.12	0.83	1.5	0.456	62	84	1.27	0.83	1.92	0.271	0.638	0.968

Table4.(Continued)

Gene	SNP Name	Comparison	Co Nev Asb	Ca Nev Asb	Nev OR	Nev CI1	Nev CI2	Nev p-value	Co Ever Asb	Ca Ever Asb	Ever OR	Ever CI1	Ever CI2	Ever p-value	LRT p-value	LRT FDR p-value
GSTM4	*rs650985*	*1) Trend*	*1311*	*1075*	*1.04*	*0.76*	*1.43*	*0.804*	*484*	*600*	*0.46*	*0.27*	*0.77*	*0.003*	*0.008*	*0.754*
GSTM4	rs650985	2) AA=0, AB=1	123	91	NA	NA	NA	NA	53	33	NA	NA	NA	NA	NA	NA
GSTM4	rs650985	3) AA=0, BB=1	3	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
GSTM4	rs650985	4) AA=0, AB+BB=1	126	94	1.02	0.73	1.42	0.908	NA	NA	0.46	0.27	0.77	0.003	0.010	0.751
GSTM4	rs650985	5) AA+AB=0, BB=1	3	3	2.30	0.34	15.52	0.394	NA	NA	2.30	0.34	15.52	0.394	NA	NA

Abbreviations: Ca, cases; CI1, lower confidence interval; CI2, upper confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; OR, odds ratio; NA, not available; NevAsb, never exposed to asbestos.
ORs calculated with unconditional logistic regression models, adjusted for the matching variables (age, sex, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.
Comparison: 1) Test for trend; 2) - 3) Additive model; 4) Dominant model; 5) Recessive model.
$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests (LRT) between the model with and without interaction term for joint exposure to the genetic variant (SNP: 0,1, 2 variant) and asbestos (never/any exposure). Both row and FDR corrected LRT p-values are reported.
Reference category: never exposed to both the genetic variant and asbestos.

Men: The nominally significant interaction with asbestos exposure among all subjects for the two SNPs tagging the GSTM4 gene was confirmed among men again in an additive model. Of note, in a recessive model the SNP rs668163 tagging the GSTA3 gene showed a borderline statistically significant interaction (FDR-corrected LRT p value = 0.102) with asbestos exposure (**Table 5**).

Table 5. ORs and 95% CIs of lung cancer for SNPs by never/ever asbestos exposure for significant (bold) or nominally significant (italics) SNP-asbestos interactions in the EAGLE study, Lombardy, Italy, 2002-2005. Men only.

Gene	**SNP Name**	**Comparison**	**Co NevAsb**	**Ca NevAsb**	**Never OR**	**Never CI1**	**Never CI2**	**Never p-value**	**Co Ever**	**Ca Ever**	**Ever OR**	**Ever CI1**	**Ever CI2**	**Ever p-value**	**LRT p-value**	**LRT FDR p-value**
up GSTM4	*rs668413*	*1) Trend*	*360*	*316*	*0.97*	*0.83*	*1.14*	*0.739*	*186*	*204*	*1.29*	*1.05*	*1.59*	*0.015*	*0.033*	*0.909*
up GSTM4	rs668413	2) AA=0, AB=1	501	398	0.88	0.69	1.12	0.307	237	277	1.08	0.79	1.49	0.624	0.097	0.888
up GSTM4	rs668413	3) AA=0, BB=1	161	124	1.00	0.71	1.40	0.995	67	110	1.83	1.19	2.84	0.006	NA	NA
up GSTM4	rs668413	4) AA=0, AB+BB=1	662	522	0.91	0.72	1.14	0.413	304	387	1.24	0.92	1.67	0.165	0.109	0.969
up GSTM4	rs668413	5) AA+AB=0, BB=1	161	124	1.08	0.79	1.46	0.644	67	110	1.75	1.18	2.61	0.006	0.056	0.768
GSTM4	*rs560018*	*1) Trend*	*432*	*376*	*0.99*	*0.84*	*1.17*	*0.885*	*230*	*246*	*1.30*	*1.05*	*1.61*	*0.015*	*0.045*	*0.909*
GSTM4	rs560018	2) AA=0, AB=1	472	361	0.88	0.7	1.11	0.288	201	263	1.40	1.03	1.91	0.031	0.056	0.888
GSTM4	rs560018	3) AA=0, BB=1	110	89	1.11	0.76	1.62	0.602	58	78	1.59	1.00	2.53	0.051	NA	NA
GSTM4	rs560018	4) AA=0, AB+BB=1	582	450	0.92	0.74	1.15	0.462	259	341	1.44	1.08	1.93	0.013	*0.016*	0.969
GSTM4	rs560018	5) AA+AB=0, BB=1	110	89	1.18	0.82	1.69	0.374	58	78	1.34	0.87	2.08	0.188	0.653	0.920
up GSTA3	*rs668163*	*1) Trend*	*411*	*355*	*0.96*	*0.82*	*1.13*	*0.601*	*213*	*249*	*1.23*	*0.99*	*1.52*	*0.060*	*0.068*	*0.909*
up GSTA3	rs668163	2) AA=0, AB=1	464	380	1.05	0.83	1.34	0.661	231	256	0.88	0.65	1.20	0.427	*0.002*	0.241
up GSTA3	rs668163	3) AA=0, BB=1	147	103	0.85	0.6	1.21	0.376	46	86	2.11	1.29	3.45	0.003	NA	NA
up GSTA3	rs668163	4) AA=0, AB+BB=1	611	483	1.00	0.8	1.25	0.973	277	342	1.05	0.79	1.4	0.743	0.810	0.984
up GSTA3	**rs668163**	**5) AA+AB=0, BB=1**	**147**	**103**	**0.83**	**0.6**	**1.15**	**0.267**	**46**	**86**	**2.24**	**1.41**	**3.58**	**0.001**	**0.001**	**0.102**
up GSTA3	rs9296695	1) Trend	858	727	0.81	0.6	1.10	0.177	413	505	1.06	0.72	1.55	0.768	0.285	0.911
up GSTA3	rs9296695	2) AA=0, AB=1	151	106	0.83	0.6	1.15	0.273	72	76	0.84	0.55	1.27	0.400	0.053	0.888
up GSTA3	rs9296695	3) AA=0, BB=1	7	3	0.44	0.08	2.4	0.342	2	7	8.53	1.41	51.77	0.020	NA	NA
up GSTA3	rs9296695	4) AA=0, AB+BB=1	158	109	0.82	0.6	1.12	0.215	74	83	0.94	0.62	1.42	0.768	0.600	0.984
up GSTA3	*rs9296695*	*5) AA+AB=0, BB=1*	*7*	*3*	*0.45*	*0.08*	*2.46*	*0.356*	*2*	*7*	*8.77*	*1.45*	*53.15*	*0.018*	*0.015*	*0.624*

Abbreviations: Ca, cases; CI1, lower confidence interval; CI2, upper confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; OR, odds ratio; NA, not available; NevAsb, never exposed to asbestos.

ORs calculated with unconditional logistic regression models, adjusted for the matching variables (age, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.

Comparison: 1) Test for trend; 2) - 3) Additive model; 4) Dominant model; 5) Recessive model.

$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests (LRT) between the model with and without interaction term for joint exposure to the genetic variant (SNP: 0,1, 2 variant) and asbestos (never/any exposure). Both row and FDR corrected LRT p-values are reported. Reference category: never exposed to both the genetic variant and asbestos.

Women: The SNP rs668413 tagging GSTM4 showed a borderline statistically significant interaction with asbestos exposure in a dominant model (FDR-LRT p-value = 0. 065), and a nominally significant interaction in the additive model (LRT p-value = 0.008). In the same gene, the SNP rs12745189 showed a statistically significant interaction with asbestos exposure in a recessive model (FDR-LRT p value = 0.013). In addition, SNPs tagging GSTM3 (rs4970774) and SOD2 (sod2_05) genes showed in an additive model a borderline interaction (FDR-p value ~10%) with asbestos exposure. It is important to underline that the few women exposed to asbestos rendered these estimates instable (**Table 6**).

Table 6. ORs and 95% CIs of lung cancer for SNPs by never/ever asbestos exposure for significant (bold) or nominally significant (italics) SNP-asbestos interactions in the EAGLE study, Lombardy, Italy, 2002-2005. Women only.

Gene	**SNP Name**	**Comparison**	**Co NevAsb**	**Ca NevAsb**	**Nev OR**	**Nev CI1**	**Nev CI2**	**Nev p-value**	**Co Ever Asb**	**Ca Ever Asb**	**Ever OR**	**Ever CI1**	**Ever CI2**	**Ever p-value**	**LRT p-value**	**LRT FDR p-value**
up GSTM4	*rs12745189*	*1) Trend*	*114*	*87*	*1.09*	*0.87*	*1.36*	*0.463*	*14*	*14*	*0.47*	*0.23*	*0.96*	*0.037*	*0.024*	0.375
up GSTM4	rs12745189	2) AA=0, AB=1	206	153	0.96	0.65	1.43	0.851	17	26	2.06	0.67	6.29	0.206	0.000	0.031
up GSTM4	rs12745189	3) AA=0, BB=1	96	92	1.19	0.76	1.88	0.446	16	2	0.09	0.01	0.58	0.011	NA	NA
up GSTM4	rs12745189	4) AA=0, AB+BB=1	302	245	1.04	0.72	1.50	0.848	33	28	0.99	0.35	2.84	0.990	0.940	0.983
up GSTM4	**rs12745189**	**5) AA+AB=0, BB=1**	**96**	**92**	**1.22**	**0.84**	**1.78**	**0.302**	**16**	**2**	**0.06**	**0.01**	**0.32**	**0.001**	**0.000**	**0.013**
up GSTM4	*rs668413*	*1) Trend*	*162*	*142*	*0.97*	*0.77*	*1.22*	*0.779*	*23*	*5*	*2.99*	*1.30*	*6.85*	*0.010*	*0.008*	*0.205*
up GSTM4	rs668413	2) AA=0, AB=1	197	138	0.80	0.56	1.14	0.218	19	30	8.71	2.48	30.61	0.001	0.001	0.075
up GSTM4	rs668413	3) AA=0, BB=1	57	52	1.04	0.64	1.71	0.867	5	7	5.13	0.9	29.15	0.065	NA	NA
up GSTM4	**rs668413**	**4) AA=0, AB+BB=1**	**254**	**190**	**0.86**	**0.61**	**1.20**	**0.362**	**24**	**37**	**7.94**	**2.33**	**27.08**	**0.001**	**0.000**	**0.065**
up GSTM4	rs668413	5) AA+AB=0, BB=1	57	52	1.18	0.74	1.86	0.485	5	7	1.19	0.28	5.08	0.816	0.990	1.000
GSTM4	rs560018	1) Trend	201	165	0.99	0.78	1.26	0.919	24	11	1.71	0.77	3.80	0.187	0.190	0.731
GSTM4	rs560018	2) AA=0, AB=1	166	131	0.95	0.67	1.36	0.795	18	25	3.40	1.18	9.78	0.023	0.065	0.583
GSTM4	rs560018	3) AA=0, BB=1	45	35	1.00	0.58	1.73	0.991	4	6	1.24	0.22	6.92	0.809	NA	
GSTM4	*rs560018*	*4) AA=0, AB+BB=1*	*211*	*166*	*0.97*	*0.70*	*1.34*	*0.836*	*22*	*31*	*2.90*	*1.04*	*8.07*	*0.041*	*0.042*	*0.454*
GSTM4	rs560018	5) AA+AB=0, BB=1	45	35	1.03	0.61	1.73	0.918	4	6	0.62	0.13	3.08	0.560	0.562	0.939
up GSTM3	***rs4970774***	***1) Trend***	***119***	***99***	***0.93***	***0.74***	***1.17***	***0.530***	***16***	***9***	***2.96***	***1.49***	***5.86***	***0.002***	***0.001***	***0.128***
up GSTM3	rs4970774	2) AA=0, AB=1	198	163	0.99	0.67	1.45	0.940	23	20	3.50	1.08	11.33	0.036	0.006	0.203
up GSTM3	rs4970774	3) AA=0, BB=1	98	69	0.85	0.53	1.37	0.509	8	13	8.71	2.22	34.28	0.002	NA	NA
up GSTM3	rs4970774	4) AA=0, AB+BB=1	296	232	0.94	0.66	1.36	0.754	31	33	4.71	1.56	14.24	0.006	0.006	0.189
up GSTM3	*rs4970774*	*5) AA+AB=0, BB=1*	*98*	*69*	*0.86*	*0.58*	*1.28*	*0.465*	*8*	*13*	*3.89*	*1.26*	*12.02*	*0.018*	*0.013*	*0.618*
up SOD2	rs4342445	1) Trend	222	194	0.96	0.73	1.25	0.739	28	18	2.33	0.92	5.9	0.075	0.066	0.545
up SOD2	rs4342445	2) AA=0, AB=1	163	117	0.87	0.62	1.24	0.451	18	23	3.16	1.19	8.42	0.021	0.035	0.454
up SOD2	rs4342445	3) AA=0, BB=1	31	21	1.07	0.56	2.06	0.843	1	1	0.53	0.03	9.92	0.670	NA	NA
up SOD2	*rs4342445*	*4) AA=0, AB+BB=1*	*194*	*138*	*0.9*	*0.65*	*1.26*	*0.551*	*19*	*24*	*2.91*	*1.10*	*7.66*	*0.031*	*0.024*	*0.350*
up SOD2	rs4342445	5) AA+AB=0, BB=1	31	21	1.13	0.59	2.14	0.713	1	1	0.31	0.02	5.56	0.425	0.403	0.939
SOD2	*rs2758331*	*1) Trend*	*114*	*97*	*0.97*	*0.77*	*1.22*	*0.792*	*10*	*15*	*0.36*	*0.17*	*0.75*	*0.007*	*0.010*	*0.226*
SOD2	rs2758331	2) AA=0, AB=1	206	155	0.83	0.56	1.22	0.344	25	23	0.41	0.13	1.23	0.111	0.033	0.454
SOD2	rs2758331	3) AA=0, BB=1	96	80	0.95	0.60	1.51	0.840	12	4	0.12	0.03	0.57	0.008	NA	NA
SOD2	rs2758331	4) AA=0, AB+BB=1	302	235	0.87	0.60	1.25	0.444	37	27	0.31	0.11	0.88	0.028	0.067	0.524
SOD2	rs2758331	5) AA+AB=0, BB=1	96	80	1.07	0.73	1.58	0.719	12	4	0.21	0.05	0.85	0.029	0.021	0.618

Table 6.(Continued)

SOD2	***sod2_05***	***1) Trend***	***103***	***90***	***0.99***	***0.79***	***1.24***	***0.911***	***13***	***6***	***3.47***	***1.66***	***7.28***	***0.001***	***0.001***	***0.128***
SOD2	sod2_05	2) AA=0, AB=1	216	160	0.84	0.56	1.25	0.384	29	23	3.04	0.82	11.24	0.096	0.005	0.203
SOD2	sod2_05	3) AA=0, BB=1	113	91	0.97	0.62	1.53	0.900	8	16	11.72	2.63	52.12	0.001	NA	NA
SOD2	sod2_05	4) AA=0, AB+BB=1	329	251	0.88	0.61	1.28	0.509	37	39	4.59	1.3	16.18	0.018	0.011	0.267
SOD2	sod2_05	5) AA+AB=0, BB=1	113	91	1.09	0.76	1.58	0.639	8	16	5.08	1.7	15.18	0.004	0.008	0.618

Abbreviations: Ca, cases; CI1, lower confidence interval; CI2, upper confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; OR, odds ratio; NA, not available; NevAsb, never exposed to asbestos.

ORs calculated with unconditional logistic regression models, adjusted for the matching variables (age, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.

Comparison: 1) Test for trend; 2) - 3) Additive model; 4) Dominant model; 5) Recessive model.

$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests (LRT) between the model with and without interaction term for joint exposure to the genetic variant (SNP: 0,1, 2 variant) and asbestos (never/any exposure). Both row and FDR corrected LRT p-values are reported. Reference category: never exposed to both the genetic variant and asbestos.

Analyses by the major histology types: Restricted to adenocarcinoma, squamous carcinoma and small cell carcinoma cases a few SNP-asbestos interactions were found. After the correction with FDR method none of them was confirmed. Of note, among adenocarcinoma cases, the most frequent histology in our study base, using an additive model the best LRT p-value between ever and never exposed to asbestos for p_{trend} values was again for the SNP rs650985 tagging the GSTM4 gene (LRT p-value = 0.005) (Data not shown).

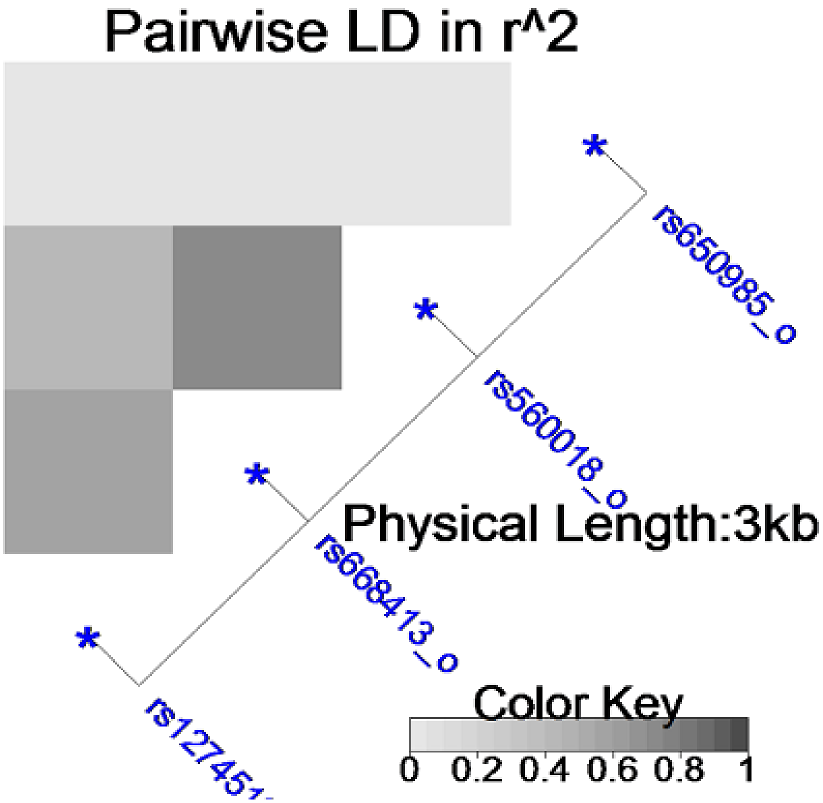
SNP grouped by genes analysis

In the grouped SNP analyses to increase the detection power I excluded the SNPs in strong LD as stated in the Methods. I measured the association between the allele pairs as r^2 correlation coefficients and represented them on a genetic map using the LDheatmap R-package (**Graph 2**). A few SNPs were eliminated from the analysis. **Table 7** shows the paired LD among the four SNPs covering the GSTM4 gene as an example: no strong LD (r^2<0.80) has resulted.

Table 7. Paired LD in r^2 among the four SNPs tagging the GSTM4 gene.

	rs12745189	rs668413	rs560018	rs650985
rs12745189	**1.00**	**0.57**	**0.42**	**0.04**
rs668413		**1.00**	**0.74**	**0.03**
rs560018			**1.00**	**0.02**
rs650985				**1.00**

Graph 2. Genetic map showing the association on a grey colour scale between the four SNPs tagging the GSTM4 gene.



SNP grouped by genes “cumulative” analysis

Among all subjects I confirmed the interaction nominally significant between the SNPs tagging GSTM4 gene and asbestos exposure (LRT p-value =0.006). Also, I found a cumulative effect with a positive trend for number of variants within the gene (p_{trend} value=0.014) that did not remain significant after the FDR correction (**Table 8**).

Table 8. ORs and 95%CIs of lung cancer for grouped SNP “cumulative” effect tagging GSTM4 gene by never/ever asbestos exposure and LRT p –values of interaction, in the EAGLE study, Lombardy, Italy, 2002-2005. All subjects.

	Co	**Ca**	**OR**	**CI1**	**CI2**	**p-val**	**Co Nev**	**Ca Nev**	**OR Nev**	**CI1 Nev**	**CI2 Nev**	**p-val Nev**	**Co Ever**	**Ca Ever**	**OR Ever**	**CI1 Ever**	**CI2 Ever**	**p-val Ever**	**LRT p-value**	**LRT FDR p-value**
GSTM4 N. SNPs																				
0	19	14	1.09	0.995	1.184	0.063	12	9	1.00	0.897	1.109	0.994	7	4	1.31	1.113	1.532	0.002		
1	220	196	1.10	0.479	2.535	0.819	159	130	1.15	0.412	3.206	0.791	60	50	0.94	0.215	4.123	0.936		
2	770	757	1.31	0.581	2.932	0.518	548	476	1.30	0.477	3.523	0.611	220	236	1.32	0.316	5.481	0.705	**0.006**	0.137
3	737	701	1.29	0.575	2.903	0.535	551	410	1.07	0.394	2.910	0.894	185	251	1.94	0.465	8.072	0.364		
4	217	230	1.56	0.679	3.570	0.296	155	130	1.38	0.495	3.856	0.538	62	86	1.94	0.450	8.387	0.373		
P trend																			***0.014***	*0.347*

Abbreviations: Ca, cases; CI1, lower confidence interval; CI2, upper confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; OR, odds ratio; N., number; Nev, never exposed to asbestos.
ORs calculated with unconditional logistic regression models, adjusted for the matching variables (age, sex, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.
$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests between the model with and without interaction term for joint exposure to the genetic variant (SNP: 0,1, 2 variant) and asbestos (never/any exposure). P $_{\text{trend}}$ for number of genetic variants within gene. Both row and FDR corrected p-values are reported.
Reference category: never exposed to both the genetic variant and asbestos.

Not significant association were found in the subgroup analyses by gender and histology (data not shown).

SNP grouped by genes “score” analysis

Among all subjects I confirmed the interaction between the SNP group tagging GSTM4 gene and asbestos exposure. In specific, the score, that I created to take into account both the number and the effect (expressed by the regression coefficient β) of each variant within the gene, resulted significantly associated with asbestos for lung cancer risk even after the FDR correction (p-value <0.000) (**Table 9**).

Table 9. ORs and 95%CIs of lung cancer for grouped SNP "score" for all 23 genes by never/ever asbestos exposure and corresponding LRT p-values of interaction, in the EAGLE study, Lombardy, Italy, 2002-2005. All subjects.

SNP Group	N SNP	Co Nev	Ca Nev	Co Ever	Ca Ever	LRT	LRT	LRT	LRT	LRT
		Asb	Asb	Asb	Asb		Nev Asb	Ever Asb	P- value	FDR p-value
GSTM4	**4**	**1425**	**1155**	**534**	**627**	**0.257**	**0.930**	**<0.001**	**<0.001**	**<0.001**
GSTM2	3	1436	1169	537	633	0.746	0.359	0.402	0.154	0.600
GSTM5	5	1426	1161	533	630	0.694	0.759	0.425	0.497	0.761
GSTM3	13	1372	1104	519	594	0.119	0.668	0.314	0.384	0.679
UGT1A7	39	1295	1050	492	565	0.175	0.127	0.750	0.196	0.600
ABCG2	20	1414	1147	530	622	0.958	0.726	0.686	0.853	0.902
GSTCD	18	1394	1148	524	617	0.272	0.166	0.755	0.382	0.679
GSTA2	3	1427	1160	532	626	0.775	0.865	0.145	0.081	0.600
GSTA1	4	1389	1127	523	612	0.017	0.241	0.024	0.273	0.600
GSTA3	7	1423	1163	532	628	0.566	0.821	0.115	0.287	0.600
GSTA4	16	1398	1127	524	610	0.212	0.522	0.076	0.863	0.902
SOD2	5	1421	1143	533	621	0.002	0.175	0.011	0.141	0.600
MDR1	33	1389	1138	522	621	0.213	0.490	0.031	0.277	0.600
NAT1	11	1424	1156	533	627	0.385	0.814	0.115	0.197	0.600
NAT2	15	1420	1158	531	629	0.366	0.709	0.268	0.509	0.761
CAT	24	1321	1078	495	587	0.375	0.153	0.681	0.845	0.998
GSTP1	7	1405	1145	530	620	0.566	0.195	0.444	0.297	0.902
ALDH2	7	1426	1157	531	624	0.243	0.341	0.126	0.778	0.902
GSTZ1	11	1404	1151	523	625	0.347	0.278	0.558	0.925	0.925
NQO1	8	1402	1139	529	620	0.806	0.932	0.666	0.560	0.761
MPO	7	1399	1142	527	621	0.852	0.946	0.889	0.596	0.761

Table 9. (Continued)

COMT	29	1207	1040	470	575	0.982	0.887	0.970	0.217	0.600
GSTT2	2	1420	1141	530	618	0.208	0.427	0.330	0.593	0.761

Abbreviations: Ca, cases; CI, confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; OR, odds ratio; NevAsb, never exposed to asbestos.

$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests between the logistic model adjusted for the matching variables (age, sex, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls with and without interaction term for joint exposure to the genetic variant (grouped SNP "score" effect) and asbestos (never/any exposure).

Both row and FDR corrected p-values are reported.

Reference category: never exposed to both the genetic variant and asbestos.

Not significant association were found in the subgroup analyses by gender and histology (data not shown).

Pathway analysis

Among the six pathways evaluated, the **GSTM** (p= 0.036) and **antioxidant** (p= 0.018) pathways, driven by **GSTM4** and **SOD2 genes,** respectively, resulted associated with asbestos exposure for lung cancer risk. The p values reported were adjusted for the number of SNPs within each gene, but not corrected for multiple comparisons since each pathway evaluated can be considered as an independent test (**Table 10**).

Table 10. Association between the six pathways evaluated and ever/never asbestos exposure for lung cancer risk. All subjects.

Pathway by Asbestos exposure	N. gene	N. SNP	P value	Most significant genes
Antioxidant Asbestos Never	2	29	0.173	CAT,SOD2
Antioxidant Asbestos Ever	2	29	**0.036**	**SOD2**
GST Asbestos Never	12	89	0.747	GSTA4,GSTM2,GSTM5,GSTCD,GSTA2,GSTA1,GSTA3
GST Asbestos Ever	12	89	0.099	GSTM4,GSTM3
GSTA Asbestos Never	4	29	0.454	GSTA4
GSTA Asbestos Ever	4	29	0.317	GSTA3,GSTA2,GSTA1
GSTM Asbestos Never	4	24	0.877	GSTM5,GSTM2
GSTM Asbestos Ever	4	24	**0.018**	**GSTM4**,GSTM5,GSTM3
NAT Asbestos Never	2	26	0.789	NAT2
NAT Asbestos Ever	2	26	0.479	NAT1,NAT2
ALL GENES Asbestos Never	23	293	0.906	GSTA4,UGT1A7,GSTCD,SOD2,MDR1,CAT,GSTP1
ALL GENES Asbestos Ever	23	293	0.080	GSTM4,GSTM3,SOD2,MDR1

Gene-level P-values across the candidate genes included in the selected biological pathway through an adaptive rank-truncated product (ARTP) method that uses a permutation algorithm for the evaluation of its significant level.

Haplotype analysis

Interestingly, the haplotype analysis for the 4 SNPs in GSTM4 (which were in low LD for most SNPs pairs) revealed two haplotypes with a borderline association with lung cancer in the overall population. Using the most frequent haplotype as reference (TGAA, freq= 47%), the carriers of the haplotype CGAG (freq=4%), and CTAA (freq=6%) showed a positive (OR= 1.48; p-value=0.062) and a negative (OR= 0.60; p-value=0.069) interaction with asbestos exposure, respectively (**Table 11**).

Table 11. Frequency of haplotypes estimated for GSTM4 gene and interaction effect with never/ever asbestos exposure for lung cancer risk. All subjects.

GSTM4	**Locus1**	**Locus2**	**Locus3**	**Locus4**	**Haplotype frequency**	**OR for interaction with asbestos**	**P-value interaction**
Haplotype 1	**C**	**G**	**A**	**A**	**0.09**	**0.80**	**0.224**
Haplotype 2	**C**	**G**	**A**	**G**	**0.04**	**0.60**	**0.062**
Haplotype 3	**C**	**T**	**A**	**A**	**0.06**	**1.48**	**0.069**
Haplotype 4	**C**	**T**	**G**	**A**	**0.33**	**1.23**	**0.072**
Haplotypes rare (grouped)	**-**	**-**	**-**	**-**	**0.01**	**3.48**	**0.173**
Haplotype most frequent	***T***	***G***	***A***	***A***	***0.47***	***Ref***	**-**

ORs calculated with unconditional logistic regression models, adjusted for the matching variables (age, sex, and residential area) and for tobacco smoking: cumulative exposure (pack-years), intensity (cigarettes per day), and years since quitting, categorized according to the quartiles of distribution of exposure among controls.

$P_{\text{interaction}}$ values were calculated from 2-df log-likelihood ratio tests between the model with and without interaction term for joint exposure to the haplotype variant under evaluation and asbestos exposure (never/ever).

Gene expression analysis

To follow up the previous results, I focused this analysis on GSTM4 gene. Available Affymetrix probes 210912_x_at and 204149_s_at for GSTM4 were used in the analysis. Gene expression levels from blood of controls and cases (data not shown), non-involved lung tissue cells, and tumor cells of lung cases were consistently strongly down-regulated in subjects carrying the rs12745189 variant compared to subjects with normal variant of GSTM4. On the contrary, an upper-regulation was found for subjects carrying the rs668413 variant. No difference by asbestos exposure was found. Of note, the few subjects with gene expression data and exposure to asbestos rendered the estimates unstable. This effect was particularly clear in the non-involved (**Table 12)** compared to the tumour (**Table 13)** tissue samples of the cases, likely because of the confounding effect of the high rate of chromosomal abnormalities present at the tumor level.

Table 12. Fold-Change (FC) of expression between individuals with different genetic variants among never and ever exposed to asbestos in non-involved lung tissues samples of cases, in the EAGLE study, Lombardy, Italy, 2002-2005. All subjects.

Gene Name	Affymetrix Probe Name	SNP Name	Comparison	N Subjects	FC	P-value	N Never	FC Never	P-value Never	N Ever	FC Ever	P-value Ever
GSTM4	*204149_s_at*	*rs12745189*	*1) Trend*	*13*	*0.90*	*0.002*	*12*	*0.92*	*0.028*	*1*	*0.85*	*0.087*
GSTM4	204149_s_at	rs12745189	2) AA=0, AB=1	18	0.94	0.246	11	0.94	0.327	7	1.04	0.827
GSTM4	204149_s_at	rs12745189_	3) AA=0, BB=1	8	0.79	0.001	5	0.84	0.027	3	0.81	0.246
GSTM4	204149_s_at	rs12745189	4) AA=0, AB+BB=1	26	0.89	0.039	16	0.91	0.096	10	0.96	0.844
GSTM4	210912_x_at	rs12745189	1) Trend	13	0.99	0.768	12	0.99	0.781	1	0.99	0.914
GSTM4	210912_x_at	rs12745189	2) AA=0, AB=1	18	0.97	0.474	11	0.99	0.883	7	0.83	0.094
GSTM4	210912_x_at	rs12745189	3) AA=0, BB=1	8	0.99	0.860	5	0.98	0.788	3	0.90	0.346
GSTM4	210912_x_at	rs12745189	4) AA=0, AB+BB=1	26	0.97	0.542	16	0.99	0.816	10	0.85	0.137
GSTM4	204149_s_at	rs668413	1) Trend	16	1.11	0.003	11	1.07	0.054	5	NA	NA
GSTM4	**204149_s_at**	**rs668413**	**2) AA=0, AB=1**	**17**	**1.16**	**0.006**	**11**	**1.11**	**0.102**	**6**	**1.29**	**0.009**
GSTM4	204149_s_at	rs668413	3) AA=0, BB=1	6	1.21	0.010	6	1.14	0.078	0	NA	NA
GSTM4	**204149_s_at**	**rs668413**	**4) AA=0, AB+BB=1**	**23**	**1.17**	**0.002**	**17**	**1.12**	**0.046**	**6**	**1.29**	**0.009**
GSTM4	*210912_x_at*	*rs668413*	*1) Trend*	*16*	0.99	*0.774*	*11*	1.01	0.847	*5*	NA	NA
GSTM4	210912_x_at	rs668413	2) AA=0, AB=1	17	0.94	0.175	11	0.96	0.479	6	0.91	0.117
GSTM4	210912_x_at	rs668413	3) AA=0, BB=1	6	1.01	0.825	6	1.03	0.703	0	NA	NA
GSTM4	210912_x_at	rs668413	4) AA=0, AB+BB=1	23	0.96	0.323	17	0.98	0.741	6	0.91	0.117
GSTM4	204149_s_at	rs560018	1) Trend	17	1.08	0.037	12	1.04	0.328	5	NA	NA
GSTM4	**204149_s_at**	**rs560018**	**2) AA=0, AB=1**	**17**	**1.15**	**0.014**	**11**	**1.10**	**0.159**	**6**	**1.29**	**0.009**

Table 12. (Continued)

GSTM4	204149_s_at	rs560018	3) AA=0, BB=1	5	1.12	0.159	5	1.05	0.507	0	NA	NA
GSTM4	**204149_s_at**	**rs560018**	**4) AA=0, AB+BB=1**	**22**	**1.14**	**0.012**	**16**	**1.08**	**0.172**	**6**	**1.29**	**0.009**

Abbreviations: Ca, cases; CI, confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; FC, Fold-Change (FC = 2^{β}) of expression between individuals with different genetic variants among never and ever exposed to asbestos; N, number.

Comparison: 1) Test for trend; 2) - 3) Additive model; 4) Dominant model.

Reference category: never exposed to the genetic variant. Significant and nominally significant FCs are represented in bold and italics, respectively.

Table 13. Fold-Change (FC) of expression between individuals with different genetic variants among never and ever exposed to asbestos in tumour lung tissues samples of cases, in the EAGLE study, Lombardy, Italy, 2002-2005.

Gene Name	Affymetrix Probe Name	SNP Name	Comparison	N Subjects	FC	P-value	N Never	FC Never	P-value Never	N Ever	FC Ever	P-value Ever
GSTM4	204149_s_at	rs12745189	1) Trend	15	0.91	0.174	12	0.93	0.408	3	0.86	0.151
GSTM4	204149_s_at	rs12745189	2) AA=0, AB=1	19	1.06	0.621	12	1.11	0.480	7	1.01	0.955
GSTM4	204149_s_at	rs12745189	3) AA=0, BB=1	12	0.82	0.138	9	0.86	0.341	3	0.74	0.146
GSTM4	204149_s_at	rs12745189	4) AA=0, AB+BB=1	31	0.96	0.712	21	0.99	0.966	10	0.92	0.635
GSTM4	210912_x_at	rs12745189	1) Trend	15	0.95	0.081	12	0.94	0.075	3	0.98	0.776
GSTM4	210912_x_at	rs12745189	2) AA=0, AB=1	19	0.96	0.483	12	0.97	0.661	7	1.00	0.981
GSTM4	210912_x_at	rs12745189	3) AA=0, BB=1	12	0.90	0.081	9	0.89	0.070	3	0.96	0.787
GSTM4	210912_x_at	rs12745189	4) AA=0, AB+BB=1	31	0.94	0.193	21	0.94	0.223	10	0.99	0.902
GSTM4	204149_s_at	rs668413	1) Trend	18	1.07	0.296	15	1.07	0.445	3	1.16	0.151
GSTM4	204149_s_at	rs668413	2) AA=0, AB=1	18	1.26	0.044	11	1.33	0.058	7	1.36	0.083
GSTM4	204149_s_at	rs668413	3) AA=0, BB=1	10	1.10	0.464	7	1.06	0.707	3	1.35	0.146
GSTM4	204149_s_at	rs668413	4) AA=0, AB+BB=1	28	1.20	0.076	18	1.22	0.132	10	1.36	0.060
GSTM4	210912_x_at	rs668413	1) Trend	18	1.03	0.386	15	1.04	0.252	3	1.02	0.776
GSTM4	210912_x_at	rs668413	2) AA=0, AB=1	18	1.08	0.124	11	1.13	0.028	7	1.04	0.767
GSTM4	210912_x_at	rs668413	3) AA=0, BB=1	10	1.04	0.532	7	1.05	0.455	3	1.04	0.787
GSTM4	210912_x_at	rs668413	4) AA=0, AB+BB=1	28	1.07	0.164	18	1.10	0.057	10	1.04	0.738
GSTM4	204149_s_at	rs560018	1) Trend	21	1.07	0.317	17	1.09	0.341	4	1.09	0.463
GSTM4	204149_s_at	rs560018	2) AA=0, AB=1	18	1.30	0.017	11	1.38	0.026	7	1.31	0.100
GSTM4	204149_s_at	rs560018	3) AA=0, BB=1	7	1.03	0.842	5	1.04	0.848	2	1.08	0.718
GSTM4	204149_s_at	rs560018	4) AA=0, AB+BB=1	25	1.22	0.052	16	1.26	0.075	9	1.25	0.141
GSTM4	210912_x_at	rs560018	1) Trend	21	1.02	0.603	17	1.03	0.484	4	1.02	0.772
GSTM4	210912_x_at	rs560018	2) AA=0, AB=1	18	1.08	0.116	11	1.16	0.007	7	0.99	0.903
GSTM4	210912_x_at	rs560018	3) AA=0, BB=1	7	1.00	0.956	5	0.98	0.743	2	1.06	0.709
GSTM4	210912_x_at	rs560018	4) AA=0, AB+BB=1	25	1.06	0.228	16	1.10	0.063	9	1.00	0.982
GSTM4	204149_s_at	rs650985	1) Trend	42	NA	NA	29	NA	NA	13	NA	NA
GSTM4	204149_s_at	rs650985	2) AA=0, AB=1	4	0.90	0.563	4	0.86	0.441	0	NA	NA
GSTM4	204149_s_at	rs650985	3) AA=0, BB=1	0	NA	NA	0	NA	NA	0	NA	NA
GSTM4	204149_s_at	rs650985	4) AA=0, AB+BB=1	4	0.90	0.563	4	0.86	0.441	0	NA	NA

Table 13. (Continued)

GSTM4	210912_x_at	rs650985	1) Trend	42	NA	NA	29	NA	NA	13	NA	NA
GSTM4	210912_x_at	rs650985	2) AA=0, AB=1	4	1.01	0.907	4	0.98	0.832	0	NA	NA
GSTM4	210912_x_at	rs650985	3) AA=0, BB=1	0	NA	NA	0	NA	NA	0	NA	NA
GSTM4	210912_x_at	rs650985	4) AA=0, AB+BB=1	4	1.01	0.907	4	-0.024	0.832	0	0.86	NA

Abbreviations: Ca, cases; CI, confidence interval; Co, controls; EAGLE, Environment And Genetics in Lung cancer Etiology; FC, Fold-Change (FC = $2^{ß}$) of expression between individuals with different genetic variants among never and ever exposed to asbestos.

Comparison: 1) Test for trend; 2) - 3) Additive model; 4) Dominant model.

Reference category: never exposed to the genetic variant. . Significant and nominally significant FCs are represented in bold and italics, respectively.

DISCUSSION

In a large population-based case-control study, I found with a candidate gene approach that polymorphisms of the phase II metabolic GSTM4 gene may play a role in lung cancer susceptibility in association with asbestos exposure. In particular, on the additive model, the SNP rs668413 showed consistently across different levels of analysis, the strongest interaction with asbestos exposure. A possible role of the polymorphisms of SOD2 and GSTM3 genes has been found among women only, but the small number of subjects exposed to asbestos rendered the estimates unreliable.

To the best of my knowledge, the GSTM4 gene has never been reported before in association with asbestos exposure and lung cancer risk. Another polymorphism of GSTM4 (rs506008) has been reported previously, but in association with lung cancer only (22).

Among the GST family genes, the null variants of GSTM1 and GSTT1 have been more frequently evaluated in association with asbestos exposure for lung cancer risk and other chronic asbestos-related lung diseases**,** but with inconsistent results (37, 39, 57-59).

Interestingly, the SNP rs668413 of the GSTM4 gene that showed the strongest interaction with asbestos exposure for lung cancer risk is in LD (r^2 =0.74) with the SNP rs560018 of the GSTM4 gene, which showed a slightly weaker interaction in my dataset. The SNP rs560018 was recently found in association with lung cancer survival as a predictor of cisplatin chemotherapy response (71), so this result seems to suggest an important functional role of both these GSTM4 polymorphisms in the progression of lung cancer. This opens up an interesting hypothesis about the underlying biological mechanism between GSTM4 polymorphisms and asbestos in the pathogenesis of lung cancer, and may even suggest a possible target for future molecular diagnostic tests and genetic therapy.

The interaction of the GSTM4 gene with asbestos exposure is biologically plausible since this gene encodes a soluble cytoplasmic glutathione S-transferases of the μ class involved in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione. The genes encoding the μ class of enzymes are organized in a gene cluster on chromosome 1p13.3 and are known to be highly polymorphic. These genetic variations can change an individual's susceptibility to carcinogens and toxins as well as affecting the toxicity and efficacy of certain drugs.

This study has several strengths. Incident lung cancer cases and randomly sampled population controls allowed the impact of occupational exposures as PAF to be estimated at the community level. The large sample size with elevated participation rates gave adequate power to detect the main genetic effects and gene-asbestos interaction effects and to perform stratified analyses by gender and histology. Detailed information on occupational history and smoking exposure was collected face-to-face by trained interviewers. The occupational exposure assessment was highly robust, as detailed lifetime job histories were codified into ISCO codes and translated into levels of carcinogen exposure by blindly applying to case status a highly reliable JEM, thereby eliminating from this study the potential for differential misclassification of exposure to occupational carcinogens. This contrasts with a recent GWAS study (42) in which Wei *et al.* failed to find a significant gene-asbestos interaction for lung cancer, most likely because their asbestos exposure assessment relied entirely on self-reporting, leading to a significant issue of recall bias. My study had excellent genotyping, confirmed by the small number of subjects with <90% call rate, and a good gene coverage by the selected tagging SNPs. Another weakness of the study of Wei *et al.* was its low GWAS chip coverage, but by using a candidate gene approach, I was able to integrate the sensitivity of common GWAS chips with the specificity of customized TaqMan probes to achieve the best coverage of the 23 selected genes. In fact, commercial SNP chips capture most, though not all, common variations in the genome, and it could partly explain why of >900 GWASs published to date, very few reported significant gene-environment interactions (72). I employed a systematic, multi-level analysis (at SNP, gene, haplotype, and pathway levels) that takes into account the biological complexity of genetic networks. I also employed an integrative analysis to evaluate the correlation between the genetic variants found associated with asbestos exposure for lung cancer risk and genetic functional variants at lung tissue level using gene expression data, although the small number of subjects with both gene expression data and asbestos exposure prevented me from finding any significant “signal” by asbestos exposure. A further strength of my study was the low exposure levels for occupational carcinogens in our study base, which is to be expected in a population-based study and which represents an ideal setting for testing gene-occupational carcinogens interaction. In fact, higher exposure levels could have masked the expected small effect (1.1–1.5-fold) of common genetic variants. Finally, as stated previously, the homogeneous nature of the study base’s genetic background means there is minimal possibility of confounding by population stratification.

This study also has some limitations. The low prevalence of exposure to occupational carcinogens among women prevented me from obtaining reliable risk estimates for them. No quantitative data for the occupational carcinogens selected were available, an inherent limitation of a large retrospective study such as this one, which means a semi-quantitative JEM is the best tool that can be applied. Residual confounding for smoking is always possible in a lung cancer study, although this was largely mitigated by carefully adjusting for smoking exposure in all of this study’s analyses. Perhaps the most important weakness of this study, and one that is shared by almost all previous candidate gene studies (56), is the lack of replication of my findings due to the unavailability of databases of comparable sample size and quality of asbestos exposure assessment.

CONCLUSIONS

In a large population-based study, I have found an interaction between occupational exposure to asbestos and GSTM4 polymorphisms, in particular the SNP rs668413, in relation to lung cancer susceptibility. This finding has never previously been reported and should be validated in further studies.

Considering my estimation that 18% of incident lung cancers (corresponding to ~800 cases) among men in Lombardy in 2005 were attributable to occupational exposure to asbestos, it is clear that we could have achieved an important goal for public health prevention had we been able to identify more susceptible subgroups and prevent them from being exposed.

Furthermore, these results provide greater understanding of the role of the GST family enzymes, in particular in relation to asbestos exposure, and call for further research into the mechanisms underlying the observed differences. GSTM4 polymorphisms should be further evaluated as potential targets of molecular diagnostic tests and therapeutic strategies for asbestos-related lung cancer. In particular, an important impact of these findings for occupational health could be screening interventions focused on susceptible workers, and recognition of and compensation for occupational lung cancers that have so far proved impossible to differentiate from those that are tobacco-related.

REFERENCES

1. American Cancer Society. Cancer Prevention & Early Detection Facts & Figures. Atlanta, GA: American Cancer Society; 2007. (http://www.cancer.org/downloads/STT/Global_Cancer_Facts_and_Figures_2007_rev.pdf).
2. Ferlay J, Autier P, Boniol M, *et al*. Estimates of the cancer incidence and mortality in Europe in 2006. Ann Oncol 2007 Mar; 18(3):581-92.
3. AIRT Working Group. Italian cancer figures. Report 2006: 1. Incidence, mortality and estimates. Epidemiol Prev 2006; 30(1 Suppl 2):8-147.
4. Sun S, Shiller JH, Gazdar AF. Lung cancer in never smokers: a different disease. Nature Publishing Group 2007; 7:778-90.
5. IARC Monographs on the Evaluation of Carcinogenic Risksto Humans. Lyon, France: International Agency for Research on Cancer; (http://monographs.iarc.fr/ENG/Classification/index.php). [Accessed January 03, 2010].
6. Siemiatycki J, Richardson L, Straif K, *et al*. Listing occupational carcinogens. Environ Health Perspect 2004; 112:1447-59.
7. Nelson DI, Concha-Barrientos M, Driscoll T, *et al*. The global burden of selected occupational diseases and injury risks: Methodology and summary. Am J Ind Med 2005; 48(6):400-418.
8. Driscoll T, Nelson DI, Steenland K, *et al*. The global burden of disease due to occupational carcinogens. Am J Ind Med 2005; 48(6):419-431.
9. De Matteis S, Consonni D, Bertazzi PA. Exposure to occupational carcinogens and lung cancer risk. Evolution of epidemiological estimates of attributable fraction. Acta Biomed 2008; 79(Suppl 1):34-42.

10. Kauppinen T, Toikkanen J, Pedersen D, *et al*. Occupational exposure to carcinogens in the European Union. Occup Environ Med 2000;57(1):10-18.

11. Mirabelli D. Estimated number of workers exposed to carcinogens in Italy, within the context of the European study CAREX (Italian). Epidemiol Prev 1999;23(4):346-359.

12. Mirabelli D, Kauppinen T. Occupational exposures to carcinogens in Italy: an update of CAREX database. Int J Occup Environ Health 2005;11(1):53-63.

13. Simonato L, Agudo A, Ahrens W, *et al*. Lung cancer and cigarette smoking in Europe: an update of risk estimates and an assessment of inter-country heterogeneity. Int J Cancer 2001; 91: 876-87.

14. Wright GS, Gruidl ME. Early detection and prevention of lung cancer. Curr Opin Oncol 2000;12:143-8.

15. Wood ME, Kelly K, Mullineaux LG, Bunn PA, Jr. The inherited nature of lung cancer: a pilot study. Lung Cancer 2000;30: 135-44.

16. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. Br J Cancer 2005; 93: 825-33.

17. Whittemore AS, Nelson LM. Study design in genetic epidemiology: theoretical and practical considerations. J Natl Cancer Inst Monogr 1999;(26):61-9.

18. Khoury MJ, Cohen BH, Beaty TB. Fundamentals of Genetic Epidemiology. Ed. Oxford University Press, New York; 1993; pp. 151-163.

19. Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990;7:177-85.

20. Bartsch H, Hietanen E. The role of individual susceptibility in cancer burden related to environmental exposure. Environ Health Perspect 1996 May;104 Suppl3:569-77.

21. Caporaso N, Goldstein A. Cancer genes: single and susceptibility: exposing the difference. Pharmacogenetics 1995; 5: 59-63.

22. Liloglou T, Walters M, Maloney P, *et al*. A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer. Lung Cancer 2002 Aug;37:143-6.

23. Kiyohara C, Yoshimasu K, Shirakawa T, Hopkin JM. Genetic polymorphisms and environmental risk of lung cancer: a review. Rev Environ Health 2004 Jan-Mar;19(1):15-38.

24. Liu G, Zhou W, Wang LI, *et al*. MPO and SOD2 polymorphisms, gender, and the risk of non-small cell lung carcinoma. Cancer Lett 2004 Oct 8;214(1):69-79.

25. Schneider J, Bernges U, Philipp M, Woitowitz HJ. GSTM1, GSTT1, and GSTP1 polymorphism and lung cancer risk in relation to tobacco smoking. Cancer Lett. 2004 May 10;208(1):65-74.

26. Wenzlaff AS, Cote ML, Bock CH, *et al*. GSTM1, GSTT1 and GSTP1 polymorphisms, environmental tobacco smoke exposure and risk of lung cancer among never smokers: a population-based study. Carcinogenesis. 2005 Feb;26(2):395-401. Epub 2004 Nov 4. Erratum in: Carcinogenesis. 2005 Apr;26(4):865.

27. Ye Z, Song H, Higgins JP, *et al*. Five glutathione s-transferase gene variants in 23,452 cases of lung cancer and 30,397 controls: meta-analysis of 130 studies. PLoS Med. 2006 Apr;3(4):e91.

28. Zhang JY, Wang Y, Prakash C. Xenobiotic-metabolizing enzymes in human lung. Curr Drug Metab. 2006 Dec;7(8):939-48.

29. Schwartz AG, Prysak GM, Bock CH, Cote ML. The molecular epidemiology of lung cancer. Carcinogenesis 2007 Mar;28(3):507-18.

30. Zienolddiny S, Campa D, Lind H, *et al*. A comprehensive analysis of phase I and phase II metabolism gene polymorphisms and risk of non-small cell lung cancer in smokers. Carcinogenesis 2008 Jun;29(6):1164-9.

31. Rotunno, M., Yu, K., Lubin, J.H., *et al*. Phase I metabolic genes and risk of lung cancer:multiple polymorphisms and mRNA expression. PLoS One 2009; 4, e5652.

32. Wang S, Wang F, Shi X, *et al*. Association between manganese superoxide dismutase (MnSOD) Val-9Ala polymorphism and cancer risk - A meta-analysis. Eur J Cancer 2009;45:2874-81.

33. Gallagher CJ, Ahn K, Knipe AL, *et al*. Association between haplotypes of manganese superoxide dismutase (SOD2), smoking, and lung cancer risk. Free Radic Biol Med 2009;46(1):20-4.

34. Anttila S, Raunio H, Hakkola J. Cytochrome P450-mediated pulmonary metabolism of carcinogens: regulation and cross-talk in lung carcinogenesis. Am J Respir Cell Mol Biol. 2011 May;44(5):583-90.

35. Rotunno M, Lam TK, Vogt A, *et al*. GSTM1 and GSTT1 copy numbers and mRNA expression in lung cancer. Mol Carcinog 2012;51 Suppl 1:E142-50.

36. Kiyohara C, Yoshimasu K, Takayama K, Nakanishi Y. Lung cancer susceptibility: are we on our way to identifying a high-risk group? Future Oncol 2007;3:617-27.

37. Anttila S, Luostarinen L, Hirvonen A, *et al*. Pulmonary expression of glutathione S-transferase M3 in lung cancer patients: association with GSTM1 polymorphism, smoking, and asbestos exposure. Cancer Res. 1995;55:3305-9.

38. Saarikoski S.T., M. Reinikainen S., Anttila A., *et al*. Role of NAT2 deficiency in susceptibility to lung cancer among asbestos-exposed individuals, Pharmacogenetics 2000; 10:183–185.

39. Stucker I., Boffetta P., Antilla S., *et al*. Lack of interaction between asbestos exposure and glutathione S-transferase M1 and T1 genotypes in lung carcinogenesis, Cancer Epidemiol. Biomarkers Prev 2001; 10: 1253–1258.

40. Schabath M.B., Spitz M.R., Delclos G.L., *et al*. Association between asbestos exposure, cigarette smoking, myeloperoxidase (MPO) genotypes, and lung cancer risk, Am J Ind Med 2002; 42:29–37.

41. Wang L.I., Neuberg D., Christiani D.C. Asbestos exposure, manganese superoxide dismutase (MnSOD) genotype, and lung cancer risk. J Occup Environ Med 2004; 46:556–564.

42. Wei S, Wang LE, McHugh MK, *et al*. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. Carcinogenesis. 2012;33:1531-7.

43. Van Damme K, Casteleyn L, Heseltine E, *et al*. Individual susceptibility and prevention of occupational diseases: scientific and ethical issues. J Occup Environ Med 1995;37:91-9.

44. Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. Lancet 2001; 357: 709-12.

45. Bertazzi PA and Mutti A. Biomarkers, disease mechanisms and role in regulatory decisions. In: Wild C, Vineis P, Garte S. Molecular Epidemiology of Chronic Diseases. Ed. Wiley-Blackwell, Chicester, 2008; pp.243-254.

46. Landi MT, Consonni D, Rotunno M, *et al*. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. BMC Public Health 2008;8:203.

47. International Standard Industrial Classification of all economic activities (ISIC). United Nations Publications ST/STAT/M.4/Rev.2/Add.1, Sales No.: E.71.XVII.8. New York: Publishing Service United Nations, 1971.

48. International Standard Classification of Occupations (ISCO, Revised 1968). Geneva, Switzerland: International Labor Office (ILO), ILO Publications, 1968, 2nd ed.

49. Goldberg M, Hémon D. Occupational epidemiology and assessment of exposure. Int J Epidemiol. 1993;22 Suppl 2:S5-9.

50. Bouyer J, Hémon D. Retrospective evaluation of occupational exposures in population-based case-control studies: general overview with special attention to job exposure matrices. Int J Epidemiol 1993;22 Suppl 2:S57-64

51. McGuire V, Nelson LM, Koepsell TD, *et al*. Assessment of Occupational Exposures in community-based case-control studies. Ann Rev Public Health 1998; 19:35-53.

52. Bouyer J, Hémon D. Studying the performance of a job exposure matrix. Int J Epidemiol. 1993;22 Suppl 2:S65-71.

53. Goldberg M, Kromhout H, Guenel P, *et al*. Job exposure matrices in industry. Int J Epidemiol 1993;22 Suppl 2: S10-5.

54. Kromhout H, Vermeulen R. Application of job-exposure matrices in studies of the general population: some clues to their performance. Eur Respir Rev 2001; 11: 80-90.

55. Peters S, Vermeulen R, Cassidy A *et al*. Comparison of exposure assessment methods for occupational carcinogens in a multi-centre lung cancer case–control study. Occup Environ Med 2010; 68:148–53.

56. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 2002; 5:391-7.

57. Nelson HH, Kelsey KT. The molecular epidemiology of asbestos and tobacco in lung cancer. Oncogene. 2002;21:7284-8.

58. Nymark P, Wikman H, Hienonen-Kempas T, Anttila S. Molecular and genetic changes in asbestos-related lung cancer. Cancer Lett. 2008;265:1-15.

59. Neri M, Ugolini D, Dianzani I, *et al*. Genetic susceptibility to malignant pleural mesothelioma and other asbestos-associated diseases. Mutat Res. 2008;659:126-36.

60 Bruzzi P, Green SB, Byar DP, *et al*. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 1985;122:904-14.

61. Wacholder S, Benichou J, Heineman EF, Hartge P, Hoover RN. Attributable risk: advantages of a broad definition of exposure. Am J Epidemiol 1994;140:303-9.

62. StataCorp. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP, 2009.

63. Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. Biometrics 1993;49:865-72.

64. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B; 1995; 57:289–300.

65. Vallyathan V, Shi X. The role of oxygen free radicals in occupational and environmental lung diseases. Environ Health Perspect. 1997;105 Suppl 1:165-77.

66. Ruosaari S, Hienonen-Kempas T, Puustinen A, *et al*. Pathways affected by asbestos exposure in normal and tumour tissue of lung cancer patients. BMC Med Genomics. 2008;1:55-64.

67. Heintz NH, Janssen-Heininger YM, Mossman BT. Asbestos, lung cancers, and mesotheliomas: from molecular approaches to targeting tumor survival pathways. Am J Respir Cell Mol Biol. 2010;42:133-9.

68. Liu G, Beri R, Mueller A, Kamp DW. Molecular mechanisms of asbestos-induced lung epithelial cell apoptosis. Chem Biol Interact. 2010;188:309-18.

69. Yu K, Li Q, Bergen AW, Pfeiffer RM, *et al*. Pathway analysis by adaptive combination of P-values. Genet Epidemiol 2009;33:700-9.

70. De Matteis S, Consonni D, Lubin JH, *et al*. Impact of occupational carcinogens on lung cancer risk in a general population. Int J Epidemiol 2012; 41: 711-21.

71. Moyer AM, Sun Z, Batzler AJ, *et al*. Glutathione pathway genetic polymorphisms and lung cancer survival after platinum-based chemotherapy. Cancer Epidemiol Biomarkers Prev. 2010;19:811-21.

72. Hindorff LA, Sethupathy P, Junkins H, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362-7.