Effect of Occupational Exposures on Lung Cancer Susceptibility: A Study of Gene-Environment Interaction Analysis

Jyoti Malhotra[1], Samantha Sartori[1], Paul Brennan[2], David Zaridze[3], Neonila Szeszenia-Dabrowska[4], Beata Świątkowska[4], Peter Rudnai[5], Jolanta Lissowska[6], Eleonora Fabianova[7], Dana Mates[8], Vladimir Bencko[9], Valerie Gaborieau[2], Isabelle Stücker[10], Lenka Foretova[11], Vladimir Janout[12], Paolo Boffetta[1]

[1]Icahn School of Medicine at Mount Sinai, New York, USA
[2]International Agency for Research on Cancer, Lyon, France
[3]Russian Cancer Research Center, Moscow, Russia
[4]Department of Epidemiology, The Nofer Institute of Occupational Medicine, Lodz, Poland
[5]National Institute of Environmental Health, Budapest, Hungary
[6]M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland
[7]Department of Occupational Health, Specialized State Health Institute, Banska Bystrica, Slovakia
[8]National Institute of Public Health, Bucharest, Romania
[9]Institute of Hygiene and Epidemiology, Charles University, First Faculty of Medicine, Prague, Czech Republic
[10]Centre for Research in Epidemiology and Population Health, INSERM, France
[11]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute and MF MU Brno, Czech Republic
[12]Department of Preventive Medicine, Faculty of Medicine, Palacky University, Olomouc, Czech Republic

Running title: Gene-occupation interactions in Lung cancer risk
Key words: Lung cancer, GWAS, Gene-environment interaction, Occupational exposures

Financial support: Subject recruitment for INCO-Copernicus study was supported by a grant from the European Commission's INCO-COPERNICUS program (Contract No IC15-CT96-0313). This project was also partially supported by a grant of the French National Cancer Institute (InCA) to P. Boffetta and MH CZ – DRO (MMCI, 00209805) to L. Foretova.

Corresponding author:
Jyoti Malhotra, MD, MPH
Division of Hematology and Oncology

Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai
One Gustave L. Levy Place, Box 1079
New York, NY, 10029
Tel: 212-241-4705, Fax: 212-241-2684
Email: jyoti.malhotra@mountsinai.org

Disclosures: The authors have declared no conflicts of interest.

Word count: 4002; Abstract: 249; Tables: 4; Figures: 1; Supplementary tables: 6

Abstract

Background: Occupational exposures are known risk factors for lung cancer. Role of genetically determined host factors in occupational exposure-related lung cancer is unclear.

Methods: We used genome-wide association (GWA) data from a case-control study conducted in six European countries from 1998-2002 to identify gene-occupation interactions and related pathways for lung cancer risk. GWA analysis was performed for each exposure using logistic regression and interaction term for genotypes and exposure was included in this model. Both SNP-based and gene-based interaction p-values were calculated. Pathway analysis was performed using three complementary methods and analyses were adjusted for multiple comparisons. We analyzed 312,605 SNPs and occupational exposure to 70 agents from 1802 lung cancer cases and 1725 cancer-free controls.

Results: Mean age of study participants was 60.1±9.1 years and 75% were male. Largest number of significant associations (p-value $\leq 1 \times 10^{-5}$) at SNP-level was demonstrated for nickel, brick dust, concrete dust and cement dust and, for brick dust and cement dust at the gene-level (p-value $\leq 1 \times 10^{-4}$). Approximately 14 occupational exposures showed significant gene-occupation interactions with pathways related to response to environmental information processing via signal transduction ($p<0.001$, FDR<0.05). Other pathways that showed significant enrichment were related to immune processes and xenobiotic metabolism.

Conclusion: Our findings suggest that pathways related to signal transduction, immune process and xenobiotic metabolism may be involved in occupational exposure-related lung carcinogenesis.

Impact: Our study exemplifies an integrative approach using pathway-based analysis to demonstrate the role of genetic variants in occupational exposure-related lung cancer susceptibility.

Introduction

Lung cancer is the most common cancer worldwide; with an estimated 1,600,000 new cases and 1,380,000 deaths annually (1, 2). The most important risk factor for lung cancer is tobacco smoking; with over 90% of lung cancer in men attributed to smoking (3, 4). In addition, environmental and occupational risk factors also contribute to the burden of lung cancer. The attributable fraction for lung cancer from occupational exposures has been reported to be as high as 7.9-16.5% in men and 1.4-4.5% in women (5-7) with asbestos, diesel engine emissions and other mixtures of polycyclic aromatic hydrocarbons, crystalline silica, arsenic and some heavy metals being some of the major contributors (8). The mechanism by which occupational exposures contribute to increased lung cancer risk is not well understood but the suggested mechanisms, which likely differ between carcinogenic agents, include DNA damage, chronic increase in inflammatory cytokines or growth factors, or due to impairment in DNA repair (7).

There is evidence to suggest that inherited genetic factors influence the development and progression of lung cancer. A family history of lung cancer in a first-degree relative is associated with a significantly increased risk of lung cancer (relative risk [RR], 1.95) (7). This association is stronger in never-smokers than in current smokers. Recent genome-wide association (GWA) studies of lung cancer have shown variations at 15q24-25.1 as a determinant of increased lung cancer risk (9-11). A second lung cancer locus has been identified through the GWA studies at

5p and includes the genes encoding TERT and CLMPTL1. In addition to these loci, there is evidence implicating loci at 6p, 13q and 22q as a risk factor for lung cancer (12-14).

Although GWA studies have been successful in identifying genes or loci associated with disease, the identified variants only explain a small proportion of the overall genetic variance (15). Gene-environment interactions may account for the missing heritability of the complex diseases (16). Genetically determined host factors may also play an important role in the pathogenesis of lung cancer associated with exposure to occupational risk factors. However, data looking at the complex interplay between genetic factors and exposure to occupational agents is limited. Possible mechanisms by which occupational exposures can interact with genetic variations leading to increased lung cancer risk include carcinogen detoxification, activation of carcinogens, and the processes of DNA damage and repair. We used the data generated in the framework of a recent GWA study to identify interactions of genetic variants and occupational exposures in lung cancer risk based on the INCO-Copernicus study. As part of this analysis, we studied gene-environment (G x E) interactions at both SNP and gene level. Pathway-based approach complements single SNP analysis by examining the cumulative contribution of functionally related genes or loci with an outcome, thereby incorporating biological knowledge into the analysis (17, 18). Therefore, we also performed pathway-based analysis to provide insight into possible biological associations.

Materials and Methods

Study participants

The INCO-Copernicus study is a case control study conducted during the period 1998 to 2002 in six Central and Eastern European countries in Hungary, Poland, Slovakia, Czech Republic, Romania and Russia (19, 20). The International Agency for Cancer Research (IARC; Lyon, France), was responsible for the overall coordination of the study. The study population comprised of all newly diagnosed lung cancer cases (age < 75 years) at the participating hospitals. A comparable group of hospital-based subjects (community-based in one of the Polish centers) matched by gender and age formed the control arm of the study. Eligible subjects (cases and controls) must have resided in the study area for at least 1 year before recruitment and the interview had to be conducted within three months of diagnosis. Patients with smoking-related conditions or other cancers were not considered eligible. Case patients and control subjects underwent an identical in-person interview during which they completed a detailed questionnaire and provided blood samples. Occupational histories were collected by trained interviewers using semi-structured questionnaires. Exposure to 66 occupational agents were coded blindly of case/control status by local groups of experts, comprising industrial hygienists and occupational physicians, based on the detailed questionnaires and knowledge on working conditions in the study areas. For the purpose of this interaction analysis, ever-exposure was defined as employment for at least six months in a job entailing exposure above background level. Participants with missing occupational exposure data (n=647) and those failing quality control steps (n=404) as described below were excluded.

Genotyping and Quality control

Genotyping on 4578 subjects (1968 lung cancer cases and 2610 controls) was performed using Illumina Sentrix HumanHap300 BeadChip containing 317,139 SNPs (10, 21). As part of quality

control, pairs of individuals with an identity-by-state (IBS) value > 0.98 were considered to be indicative of a duplicate sample and from these pairs; one of the individuals was randomly removed from the data. Samples with more than 5% missing genotype were also excluded. Additionally, samples whose reported sex did not match with the inferred sex based on the heterozygosity rate from the X chromosomes were removed (22). Samples with less than 70% European ancestry as determined by STRUCTURE (23, 24) and high degree of relatedness (sib-pairs and half-sib pairs, IBS > 0.50) were also excluded. Limited extent of population stratification as confirmed by principal components analysis was confirmed and outliers were removed (25). SNPs with less than 95% completion rate or departure from Hardy–Weinberg equilibrium (HWE < 10^{7}) were also excluded.

SNP-based analysis

Single marker association analysis was performed for each exposure using an unconditional logistic regression model. The model included additive SNP effect, age, gender, smoking history, country/center, binary occupational exposure, disease risk score and interaction term for genotypes and exposure. The p-value of the interaction term was used to assess the significance of the interaction between genetic variants and occupational exposure and p-value $\leq 1 \times 10^{-5}$ was considered to be significant. The disease risk score was calculated for each exposure using history of exposure to other occupational exposures to estimate the probability of disease occurrence in the absence of the exposure (26). All analysis was performed using PLINK (27). SNPs were mapped to associated genes using reference SNP location from UCSC Genome Browser hg18 assembly with ± 20 kb gene boundaries.

Gene-based analysis

Versatile gene-based association (VEGAS) test was used to calculate gene-level interaction p-values (28). This test incorporates information from a full set of markers (or a defined subset) within a gene and accounts for linkage disequilibrium (LD) between markers by using simulations from the multivariate normal distribution. SNPs are assigned to 17,787 autosomal genes according to UCSC Genome Browser hg18 assembly with ± 50 kb gene boundaries. The gene-based test statistic is the sum of all of the chi-squared 1 degree of freedom (df) statistics within that gene which are calculated from SNP-based p values. Results obtained using this method has been shown to be equivalent to those obtained via permutation. As this was a gene-environment interaction analysis, we opted for a slightly looser p-value threshold of $\leq 1 \times 10^{-4}$, compared to Bonferroni-corrected p-value threshold of $< 2.8 \times 10^{-6}$ (0.05/17,787).

Pathway-based interaction analysis

We used three complementary pathway-analysis methods which are listed below. Pathways or gene sets were considered to be significantly enriched if the adjusted p-value or false discovery rate (FDR) was ≤ 0.05 using any one method or, unadjusted p-value ≤ 0.05 using at least two methods.

(a) Improved Gene Set Enrichment analysis (i-GSEA4GWAS) (29): In iGSEA, max statistics or −log(*P*-value) of closely spaced SNPs in a gene are used to represent the gene. The ranked gene list with corresponding representing values is utilized to calculate each gene set's enrichment score (*ES*), i.e. a Kolmogorov–Smirnov like statistics with weight 1. SNP label permutation is performed to correct for gene and gene-set variation and significance proportion-based enrichment score (SPES) is calculated. SPES emphasizes on total significance coming from high

proportion of significant genes. This method has been shown to have improved sensitivity compared to GSEA and is more appropriate for study of the combined effects of possibly modestly associated SNPs/genes. This method uses a comprehensive pathway/gene set database with pathways integrated and curated from a variety of resources including KEGG (Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database), Biocarta, GenMAPP etc. Pathways with <20 genes or >200 genes were excluded from further analysis and SNPs were mapped to genes with ± 20 kb boundaries. False discovery rate (FDR) was used for multiple testing correction with FDR ≤ 0.05 considered to be significant.

(b) Interval-based Enrichment Analysis Tool for GWAS (INRICH) (30): This method detects enriched association signals of LD-independent genomic regions within biologically relevant gene sets. To ensure that multiple, adjacent SNPs that potentially tag the same causal variant are analyzed as one independent association unit, LD-independent intervals were generated around the associated SNPs using PLINK 1.07, and enrichment test was conducted using KEGG database and Gene Ontology terms. A permutation approach is used to calculate empirical significance *P*-values for each gene-set. This method therefore adjusts for confounding factors like varying gene size, SNP density, LD that is not usually accounted for by GSEA. Multiple testing corrections are achieved via a second, nested round of permutation to assess the null distribution of the minimum empirical *P*-value across all tested gene-sets. Adjusted p-value ≤ 0.05 was considered to be significant.

(c) Systematic biological Knowledge-based mining system (KGG) (31, 32): Using the hybrid set-based test (HYST) approach, SNPs in a gene set are partitioned into LD-based blocks. For each of these blocks, extended Simes' test (GATES) (33) is used to calculate the block-based p-value for association and to mark the key SNP from which this was derived. Finally, the scaled

chi-square test is employed to combine the *n* block-based p values into a single test statistic. 880 canonical pathways (including from KEGG, BioCarta and Reactome) collected from the MsigDB database were used to annotate pathways. Gene-set p-value < 0.0001 was considered to be significant (0.05/880 = 0.000056 on calculating FDR).

Results

Patient characteristics

A total of 3527 participants (1802 lung cancer cases, 1725 controls) were included in our analysis; their baseline characteristics are reported in Table 1. The mean age of the study participants was 60.1 ± 9.1 years and there was no difference in mean age between cases and controls. Cases had higher percentage of males (77.5% vs 72.4%; p<0.0001) as compared to controls. As expected, there was a marked difference between cases and controls in smoking habits. 71.3% of lung cancer cases had non-small cell lung cancer with squamous cell cancer (42.5%) forming the predominant histological type followed by adenocarcinoma (22.6%). Percentage of cases and controls with ever exposure to each occupational agent is listed in Supplementary Table S1. Seventeen established or suspected lung carcinogens that were included in our analysis include arsenic, asbestos (amphibole and chrysotile), metals (arsenic, cadmium, chromate, chromium, nickel and their compounds), chlorinated solvents, diesel engine emissions, organic dust and silica (brick dust, cement dust, concrete dust, respirable free crystalline silica and sand). In addition, 49 other occupational agents with a possible relation to lung carcinogenesis were also included in this analysis.

Author Manuscript Published OnlineFirst on January 12, 2015; DOI: 10.1158/1055-9965.EPI-14-1143-T
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

SNP-based association analysis

After quality control, 312605 SNPs were included in our analysis. The SNPs which showed interaction p-values (GxE) < 1×10^{-5} are reported in Supplementary tables S2 and S3. Of the established/suspected lung carcinogens, the largest number of significant associations (p-value ≤ 1×10^{-5}) was demonstrated for nickel & compounds, brick dust, concrete dust and cement dust. Asbestos (and chrysotile), chromium & compounds, organic dust and sand showed association with at least 3 SNPs each. Many exposures showed multiple SNPs associated within the same loci or gene - asbestos (3 SNPs on WDR6), chrysotile (PKNOX2), nickel & compounds (CGNL1, SUSD5), live animals (SLC35F5), cement dust (CD46 and CPA3) and, sand (LBA1). No significant association was found for amphibole asbestos, arsenic & compounds, cadmium & compounds, chromate & compounds and chlorinated solvents. Of the other occupational exposures, significant SNP-based association were observed for hard wood dust (19 SNPs), soft wood dust (11 SNPs), inorganic acid mist (10 SNPs), inorganic acids mist (9 SNPs), animal viruses (7 SNPs with 4 corresponding to CNTN1 gene and 3 to NFIC gene), plastics pyrolysis fumes (13 SNPs). Of note, diesel and kerosene exposure showed a significant association with 23 SNPs within a region on chromosome 15 with 13 SNPs mapped to the gene MYO9A.

Gene-based association analysis

Genes with G x E interaction p-values ≤ 1×10^{-4} for each occupational exposure are listed in table 2 and supplementary table S4. Of the established/suspected carcinogens, we did not find any significant gene-based interaction for asbestos, arsenic & compounds, nickel & compounds, chromium & compounds, cadmium & compounds and chlorinated solvents. The largest number of significant associations was demonstrated for brick dust (AQP8, CPA4, and CPA5) and

cement dust (ANKRD56, CD46, DAZAP2, LOC57228, and RTN4R). Associations that were significant in both SNP-based and gene-based analysis were noted for gene CD46 for cement dust, LBA1 for sand and 5 genes on chromosome 15 including MYO9A (p=0.00006) for diesel & kerosene. Of the other occupational exposures, significant gene-based associations were found for stainless steel dust, soft wood dust, inorganic acid mist and plastics pyrolysis fumes.

Significant genomic regions identified and associated genes with each region are listed in supplementary table S5. Genomic regions with largest number of significant genes identified for at least one of the gene-occupation interactions are 1q23.2, 1q24.1 to 1q24.2, 6p21.31, 11q13.2, 15q23 and 19p13.3. Main categories of proteins coded for by the significant genes identified in our analyses are listed in Supplementary table S6 using PANTHER (Protein Analysis Through Evolutionary Relationships) Classification System (34, 35). A majority of the proteins coded by these genes are either involved in the transcription process (either in nucleac acid binding or as a transcripton factor), or function as a receptor or enzyme modulator.

Pathway-level interaction analysis

The most significant pathways (p-value ≤ 0.05 after adjusting for multiple comparisons) related to interaction analysis between genetic variants and each occupational exposure is reported in Tables 3 & 4 and, Figure 1. Most of the significant pathways were related to signal transduction pathways which play an integral role in environmental information processing, immune processes and metabolism of exogenous agents.

(a) Signal transduction pathways: Mitogen activated protein kinase (MAPK) pathway or ERK1-ERK2 pathway showed significant associations with exposure to chromate & compounds,

chromium & compounds and brick dust, as well as with five other potential carcinogens (coal and coke combustion fumes, diesel and kerosene, inorganic insulation dust, iron, formaldehyde). Both chromium and chromate compounds were significantly associated with Wnt pathway. In addition, animal viruses were also associated with this pathway. Chromate also showed significant association with tumor necrosis factor receptor (TNFR) pathway. Of the silica products, both brick and cement dust were also associated with TNFR1 pathway. Although arsenic & compounds showed no significant SNP-based or gene-based interaction associations, they did show significant association with FC epsilon receptor signaling pathway (adjusted p-value 0.02). Nickel & compounds showed significant enrichment for nuclear factor-KB (NFKB) and hedgehog signaling pathway. NFKB pathway was also associated with chrysotile exposure.

(b) Pathways involved in immune regulation: Chromate & compounds showed association with B cell activation pathways (p-value 0.0007) and both wood dust and soft wood dust showed association with antigen processing and presentation. Petroleum products like gasoline and, diesel and kerosene showed significant associations with T cell activation pathways. T cell pathways were also enriched for soot, inorganic insulation dust and inorganic pesticides.

(c) Xenobiotic metabolism (metabolism of potential carcinogens): Phase I or cytochrome 450 metabolism was enriched for diesel engine emissions (adjusted p 0.06) and for plastics pyrolysis fumes. In addition, phase II metabolism was enriched for chromium & compounds.

(d) Others: Other enriched pathways included those related to p53 (chromate & compounds), DNA replication (nickel & compounds, concrete dust, petroleum oil combustion fumes), apoptosis (animal feeding, hard alloys dust, asphalt-bitumen fumes), cell death (animal feeding) and DNA repair (carbon black dust).

Discussion

Knowledge about the role of interactions between genetic factors and occupational exposures in lung carcinogenesis is limited. To our knowledge, our study is the first to use GWA data to analyze gene-occupation interactions between a large range of occupational exposures and GWA-derived genetic variants in determining lung cancer susceptibility. In addition to SNP-based and gene-based analysis, we also performed pathway-based analysis in order to identify possible causal pathways involved. We found that pathways related to signal transduction (especially MAPK and Wnt), immune response and xenobiotic metabolism were significantly enriched in relation to the studied gene-occupational exposure interactions. We found significant association at SNP-based, gene-based and pathway-based analysis in our study which indicates that interaction between genetic variants or polymorphisms and occupational agents may play a role in determining an individual's susceptibility to lung cancer.

Previous GWA studies have shown multiple loci to be associated with lung cancer susceptibility especially those related to 15q, 5p and 6p with $p < 5 \times 10^{-8}$ (9-13). However, these explain only a small proportion of individual variation in lung cancer susceptibility and gene-environment interactions may account for the missing variation (15, 16). A number of studies looking at gene-environment interactions for smoking have found strong evidence of interactions with smoking (36, 37). However, hardly any studies have looked at gene-environment interactions for occupational exposures in lung carcinogenesis although occupational risk factors have a significant contribution to the burden of lung cancer (5-7). Wei et al conducted a study to determine genome-wide gene–asbestos exposure interaction on lung cancer risk at levels of SNPs, genes and pathways. Although they reported limited power to detect gene-environment

interactions at the SNP and gene levels, their pathway-based approach did show significant associations (38) in the immune function regulation-related pathways. Our study also demonstrates significant gene-environment interactions in immune system pathways for many exposures such as heavy metals and petroleum products, but not for asbestos.

Our study looked at gene-environment interactions for nearly 70 occupational exposures and has important implications as research in this area till now has been limited. We used both SNP-based and gene-based association methods as both have their own advantages and disadvantages. SNP-based method can identify significant loci with only a small number of significant SNPs but may miss detecting SNPs with weak marginal effects, but a strong joint effect. Gene-based analysis may identify these genes with more than one causative variants with marginal levels of significance that are often indistinguishable from random noise in the initial SNP-based GWAS results (28).

One limitation of gene-environment interaction studies is the need for larger sample size and more than 10,000 case-control pairs may be required to detect significant interactions (39). Of the hundreds of published GWASs, few have reported significant SNP-level gene-environment interactions (40). Current GWAS design may not provide enough statistical power to detect interactions at the single SNP level, as it mainly focuses on the main effect of a SNP. This is especially true for analysis of rare variants with weak main effects that may not be detectable in the analysis of genetic marginal effects only (41). Therefore, we included occupational exposures for which main genetic effects have not yet been elucidated also in our analysis.

As our study is exploratory and hypothesis-generating, we used less stringent threshold p-values for significance after correcting for multiple comparisons. In addition, as SNP-based or gene-based associations may miss markers that may contribute additively and incrementally to cancer risk by participating in common pathways and networks underlying disease susceptibility, we performed pathway analysis to identify potential biological causal pathways (17, 18). There is limited agreement about the optimum methods to identify top ranked pathways. Studies have compared different pathway analysis methods and have found only a modest degree of overlap between top pathways chosen by each method (42, 43). Since different pathway analysis methods can produce different results using the same data set, it is recommended to use more than one method when examining pathway associations with disease risk (42, 43). Therefore, we used three different but complimentary pathway analysis methods with different approaches to pathway analysis.

Our results suggest that polymorphisms related to pathways involved in signal transduction (especially MAPK, Wnt and TNGR), immune response and metabolism of exogenous agents may increase individual's risk to lung cancer on being exposed to certain occupational agents. While the genes involved in these pathways have been shown to be involved in lung cancer pathogenesis (44-48), data showing the association of these pathways in relation to interactions between genetic polymorphisms and occupational exposures is lacking. Our study is one of the first studies to identify this relationship and further studies in this area are needed. Also, although we only studied potential lung carcinogens in the occupational settings, our results can potentially be extrapolated to individuals with history of exposures to these agents in other settings, like at home or in the environment.

One main concern of a gene-environment study is the potential to misclassify environmental or exposure variable which can impact power and generate bias. A major strength of our study is that it was designed specifically to measure occupational exposures and detailed occupational exposure assessment was done by local expert panels using detailed semi-structured questionnaires. Another strength of our study is that we conducted analysis at the levels of SNPs, genes and pathways, thereby reducing the chance to miss weaker associations without compromising on adjustments for multiple testing. One limitation of our study is that it does not include validation of the results using an independent dataset. This is mainly because detailed occupational data is not usually collected as part of GWA studies in lung cancer and therefore, we did not have access to similar datasets to replicate our findings. Also, due to limited sample size, we were unable to perform stratified interaction analysis by level or duration of exposure, histology, gender or country. However, with 1802 lung cancer cases and 1725 controls, the sample size in our study was at least similar if not larger than major published lung cancer GWA studies. Another potential limitation of our study is that although cases and controls were matched during initial study enrollment, this may not be applicable to the genotypic analysis as subjects were excluded before performing the GWA analysis using quality control criteria. We used unconditional logistic regression for analysis which does not account for preferential sampling through the cases. However, unconditional logistic regression has been reported to be robust to gene-environment correlation (49). In addition, there is no single method that has been validated to outperform other methods for gene-environment interaction analysis (39). In addition, an important limitation of the pathway-based approach for GWA analysis is the incomplete annotation of the human genome (3232). Also, the pathway analysis methods used in

our study have not been compared with each other previously to ascertain the degree of overlap in detecting significant associations using these methods.

In conclusion, our study exemplifies an integrative approach using SNP-based, gene-based and pathway-based analysis to demonstrate that genetic variants may play an important role in determining an individual's susceptibility to lung cancer in response to occupational exposures. Our results suggest that pathways related to signal transduction, immune processes and xenobiotic metabolism may contribute to this increased susceptibility: they should be interpreted as hypothesis-generating and request validation in independent datasets. Our study also showed that although pathway-based approaches may help in elucidating potentially important associations and causal pathways for gene-environment interaction analyses.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: a cancer journal for clinicians. 2011;61:69-90.
2. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer Journal international du cancer. 2010;127:2893-917.
3. Ezzati M, Lopez AD. Estimates of global mortality attributable to smoking in 2000. Lancet. 2003;362:847-52.
4. Tyczynski JE, Bray F, Parkin DM. Lung cancer in Europe in 2000: epidemiology, prevention, and early detection. The lancet oncology. 2003;4:45-55.
5. Boffetta P, Tubiana M, Hill C, Boniol M, Aurengo A, Masse R, et al. The causes of cancer in France. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2009;20:550-5.
6. Rushton L, Hutchings S, Brown T. The burden of cancer at work: estimation as the first step to prevention. Occupational and environmental medicine. 2008;65:789-800.
7. Olsson AC, Gustavsson P, Zaridze D, Mukeriya A, Szeszenia-Dabrowska N, Rudnai P, et al. Lung cancer risk attributable to occupational exposures in a multicenter case-control study in Central and Eastern Europe. Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine. 2011;53:1262-7.
8. Driscoll T, Nelson DI, Steenland K, Leigh J, Concha-Barrientos M, Fingerhut M, et al. The global burden of disease due to occupational carcinogens. American journal of industrial medicine. 2005;48:419-31.
9. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nature genetics. 2008;40:616-22.
10. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452:633-7.
11. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008;452:638-42.
12. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nature genetics. 2008;40:1407-9.
13. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, et al. Lung cancer susceptibility locus at 5p15.33. Nature genetics. 2008;40:1404-6.
14. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nature genetics. 2011;43:792-6.
15. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747-53.
16. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456:18-21.
17. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology. 2012;8:e1002375.
18. Zhao J, Gupta S, Seielstad M, Liu J, Thalamuthu A. Pathway-based analysis using reduced gene subsets in genome-wide association studies. BMC bioinformatics. 2011;12:17.
19. Zeka A, Mannetje A, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, Lissowska J, et al. Lung cancer and occupation in nonsmokers: a multicenter case-control study in Europe. Epidemiology. 2006;17:615-23.

20. Scelo G, Constantinescu V, Csiki I, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, et al. Occupational exposure to vinyl chloride, acrylonitrile and styrene and lung cancer risk (europe). Cancer causes & control : CCC. 2004;15:445-52.
21. Hung RJ, Brennan P, Canzian F, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, et al. Large-scale investigation of base excision repair genetic polymorphisms and lung cancer risk in a multicenter study. Journal of the National Cancer Institute. 2005;97:567-76.
22. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5:1564-73.
23. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human mutation. 2009;30:69-78.
24. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945-59.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38:904-9.
26. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Statistical methods in medical research. 2009;18:67-80.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81:559-75.
28. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. American journal of human genetics. 2010;87:139-45.
29. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic acids research. 2010;38:W90-5.
30. Lee PH, O'Dushlaine C, Thomas B, Purcell SM. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 2012;28:1797-9.
31. Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. American journal of human genetics. 2012;91:478-88.
32. Gui H, Li M, Sham PC, Cherny SS. Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's Disease dataset. BMC research notes. 2011;4:386.
33. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. American journal of human genetics. 2011;88:283-93.
34. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic acids research. 2013;41:D377-86.
35. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nature protocols. 2013;8:1551-66.
36. Zhang R, Chu M, Zhao Y, Wu C, Guo H, Shi Y, et al. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. Carcinogenesis. 2014.
37. Qiu F, Yang L, Fang W, Li Y, Yang R, Yang X, et al. A functional polymorphism in the promoter of ERK5 gene interacts with tobacco smoking to increase the risk of lung cancer in Chinese populations. Mutagenesis. 2013;28:561-7.
38. Wei S, Wang LE, McHugh MK, Han Y, Xiong M, Amos CI, et al. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. Carcinogenesis. 2012;33:1531-7.

39. Thomas DC, Lewinger JP, Murcray CE, Gauderman WJ. Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome. American journal of epidemiology. 2012;175:203-7; discussion 8-9.
40. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009;106:9362-7.
41. Hein R, Beckmann L, Chang-Claude J. Sample size requirements for indirect association studies of gene-environment interactions (G x E). Genetic epidemiology. 2008;32:235-45.
42. Fehringer G, Liu G, Briollais L, Brennan P, Amos CI, Spitz MR, et al. Comparison of pathway analysis approaches using lung cancer GWAS data sets. PloS one. 2012;7:e31816.
43. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nature reviews Genetics. 2010;11:843-54.
44. Kelly RJ, Giaccone G. Lung cancer vaccines. Cancer journal. 2011;17:302-8.
45. Zhao Y, Zeng J, Zhang Y, Lu S, Zhao E, Huang Z, et al. GSTM1 polymorphism and lung cancer risk among East Asian populations: a meta-analysis. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2014.
46. Hiraoka K, Miyamoto M, Cho Y, Suzuoki M, Oshikiri T, Nakakubo Y, et al. Concurrent infiltration by CD8+ T cells and CD4+ T cells is a favourable prognostic factor in non-small-cell lung carcinoma. British journal of cancer. 2006;94:275-80.
47. Stewart DJ. Wnt signaling pathway in non-small cell lung cancer. Journal of the National Cancer Institute. 2014;106:djt356.
48. Koul HK, Pal M, Koul S. Role of p38 MAP Kinase Signal Transduction in Solid Tumors. Genes & cancer. 2013;4:342-59.
49. Sohns M, Viktorova E, Amos CI, Brennan P, Fehringer G, Gaborieau V, et al. Empirical hierarchical bayes approach to gene-environment interactions: development and application to genome-wide association studies of lung cancer in TRICL. Genetic epidemiology. 2013;37:551-9.

Table 1. Demographic characteristics of lung cancer cases and controls

	Cases (n=1802)	Controls (n=1725)	p-value
Age, mean	60.3 ± 8.7	60.0 ± 9.4	0.34
Gender			
Male	1397 (77.5%)	1248 (72.4%)	<0.001
Female	405 (22.5%)	477 (27.7%)	
Country			
Romania	90 (5.0%)	141 (8.2%)	<0.001
Hungary	270 (15.0%)	234 (13.6%)	
Poland	608 (33.8%)	634 (36.7%)	
Russia	345 (19.2%)	260 (15.1%)	
Slovakia	290 (16.1%)	153 (8.9%)	
Czech Republic	199 (11.0%)	303 (17.6%)	
Tobacco pack-years	34.9 ± 20.9	18.1 ± 20.1	
Tobacco status			<0.001
Never smoker	138 (7.7%)	588 (34.1%)	
Ex-smoker >= 2 years	359 (19.9%)	484 (28.1%)	<0.001
Current smoker	1304 (72.4%)	649 (37.6%)	
Histology			
Adenocarcinoma	408 (22.6%)	--	
Squamous	766 (42.5%)		
Large cell	45 (2.5%)		
Small cell	278 (15.4%)		
Mixed	66 (3.7%)		

Table 2. Gene-level significant occupational exposure interactions for established/suspected occupational lung carcinogens

	CHR	GENE	No. of SNPs	Start BP	End BP	OR	p-value
Metals							
Chromate & Compounds	1	C1orf101	26	242691315	242870285	121.68	0.000052
Chlorinated solvents	1	HRNR	7	150451181	150463293	56.06	0.000078
Organic dust							
Animal feeding	19	ELAVL1	14	7929456	7976529	108.84	0.000009
Silica							
Brick dust	16	AQP8	11	25135785	25147754	64.70	0.000056
	7	CPA4	25	129720229	129751250	155.95	0.000086
	7	CPA5	27	129771865	129795807	178.97	0.000042
Cement dust	4	ANKRD56	15	78035105	78038026	83.15	0.000018
	1	CD46	12	205992024	206035481	119.34	0.000095
	12	DAZAP2	10	49918892	49923834	89.70	0.000038
	12	LOC57228	12	49925399	49950469	95.21	0.000055
	22	RTN4R	21	18608937	18635816	86.81	0.000096
Concrete dust	17	ARSG	33	63766917	63928595	125.10	0.000053
Respirable free crystalline silica	5	MGC42105	8	43228083	43316709	65.89	0.000027
Sand	2	COL5A2	14	189604885	189752850	89.21	0.000051
	3	LBA1	18	36843314	36877415	116.15	0.000023

Abbreviations: CHR, chromosome; SNP, Single nucleotide polymorphism; BP, base-pair; OR, Odds ratio

Author Manuscript Published OnlineFirst on January 12, 2015; DOI: 10.1158/1055-9965.EPI-14-1143-T
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Table 3. Pathway-based analysis using occupational exposure interactions for established/suspected occupational lung carcinogens

	Pathway name	Pathway database	Pathway analysis method	p-value	Adjusted p-value	Significant genes or genomic regions/All genes
Asbestos						
Chrysolite	WNT pathway	Biocarta	IGSEA	<0.001	0.035	11/21
	NFKB pathway	REACTOME	KGG	0.00008		6/11
Metals						
Arsenic & Compounds	FC Epsilon receptor signaling pathway	KEGG	IGSEA	<0.001	0.02	28/69
	Gap junction	KEGG	IGSEA	<0.001	0.04	35/81
Chromate & Compounds	ERK1-ERK2/MAPK pathway	STKE	KGG	0.007	-	5/32
		STKE	IGSEA	0.02	0.23	10/25
	B cell Antigen receptor	STKE	KGG	0.0007		11/40
	P53 signaling pathway	KEGG	IGSEA	<0.001	0.01	23/60
	WNT signaling pathway	Superarray	IGSEA	<0.001	0.02	25/55
	Stress pathway (TNF)	Biocarta	IGSEA	0.001	0.04	10/23
Chromium & Compounds	ERK1-ERK2 pathway	STKE	IGSEA	0.006	0.24	11/25
		Biocarta	KGG	0.0009	-	8/28
	Phase II conjugation	REACTOME	KGG	0.0002		10/70
Nickel & Compounds	NFKB pathway	Reactome	KGG	0.0003		6/23
		KEGG	IGSEA	0.003	0.14	45/130
	Purine metabolism	KEGG	KGG	0.002	-	25/159
	Hedgehog signaling pathway	KEGG	IGSEA	0.005	0.14	20/53
		KEGG	KGG	0.05	-	12/56
	DNA polymerase	KEGG	IGSEA	0.004	0.06	8/23
Diesel engine emissions	Xenobiotic metabolism by C450	KEGG	IGSEA	<0.001	0.06	26/61
		KEGG	KGG	0.06	-	14/70
	Glutathione metabolism	KEGG	IGSEA	<0.001	0.01	11/30
		KEGG	KGG	0.0015	-	15/50
Organic dust						
Animal feed	Regulation of apoptosis	GO:0043065	INRICH	0.002	0.023	2
Live animals	Snare interactions in transport	KEGG	IGSEA	<0.001	0.06	13/29
		Reactome	KGG	0.01	-	7/38
Silica						
Brick dust		KEGG	INRICH	0.0002	0.03	4
	Purine metabolism	KEGG	KGG	0.02	-	25/159
	ERBB2-ERBB3 pathway	PID	KGG	0.00004		9/55
	TNFR1 pathway	Biocarta	KGG	0.0002		11/29
Cement dust	TNFR1 pathway	Biocarta	IGSEA	0.006	0.1	11/28
		Biocarta	KGG	0.0002	-	10/29
	FAS pathway	Biocarta	IGSEA	0.004	0.08	11/27
		Biocarta	KGG	0.0002	-	10/30
Concrete dust	DNA polymerase- DNA replication	KEGG	IGSEA	<0.001	0.066	8/23
Sand	Basal transcription factors	KEGG	IGSEA	0.001	0.046	10/25

Abbreviations: IGSEA, Improved Gene Set Enrichment Analysis; KGG, Knowledge-based mining system for Genome-wide Genetic studies; KEGG, Kyoto Encyclopedia of Genes and Genomes; STKE, Signal Transduction Knowledge Environment; PID, Pathway Interaction Database

Author Manuscript Published OnlineFirst on January 12, 2015; DOI: 10.1158/1055-9965.EPI-14-1143-T
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Table 4. Pathway-based analysis using occupational exposure interactions for possible occupational lung carcinogens

	Pathway name	Pathway database	Pathway analysis method	P-values
Asphalt-bitumen fumes	BAD pathway - apoptosis	Biocarta	IGSEA	<0.001
Carbon black dust	DNA repair	REACTOME	KGG	0.0007
	Glycerolipid metabolism	KEGG	IGSEA/KGG	0.001/0.01
Coal and coke combustion fumes	Ribosome	KEGG	IGSEA	0.001
	Integrin pathway	Biocarta/PID	IGSEA/KGG	0.007/0.008
	FAS signaling pathway	STKE/PID	IGSEA/KGG	0.001/0.02
	PDGF pathway	Biocarta/Biocarta	IGSEA/KGG	0.004/0.036
	MAPK-ERK pathway	Biocarta	IGSEA/KGG	0.007/0.002
	ECM receptor interaction	KEGG	IGSEA/KGG	0.002/0.02
Coal dust	TNF pathway	STKE	IGSEA	<0.001
	Ribosomal proteins	Wiki	IGSEA	<0.001
	Fas signaling pathway	STKE/PID	IGSEA/KGG	0.004/0.018
Coal tar pitch fumes	RAS pathway	GO:0007265/PID	INRICH/KGG	0.003/0.001
	P53 pathway	Biocarta	KGG	0.00005
	GPCR pathway	Wiki	IGSEA	0.001
Graphite dust	Leukocyte migration	KEGG	IGSEA/KGG	0.008/0.009
	Snare interactions in transport	KEGG/Reactome	IGSEA/KGG	0.016/0.03
	Ribosome protein	Wiki/KEGG	IGSEA/KGG	0.002/0.048
	B cell receptor pathway	KEGG/SigmaAldrich	IGSEA/KGG	0.002/0.004
Soot	Phosphatidylinositol signaling system	KEGG	IGSEA	<0.001
	Inositol phosphate metabolism	KEGG	IGSEA/KGG	0.002/0.002
	TNF pathway	STKE	IGSEA	0.001
	T cell signal transduction	STKE/Biocarta	IGSEA/KGG	0.01/0.008
Diesel and kerosene	MAPK pathway	STKE/Reactome	IGSEA/KGG	0.024/0.03
	TCR pathway	Biocarta/PID	IGSEA/KGG	0.024/0.03
	FC Epsilon signaling pathway	KEGG	IGSEA	<0.001
	GPCR pathway	Biocarta	IGSEA	<0.001
Gasoline	TCR pathway	Biocarta/STKE	IGSEA/KGG	0.013/0.057
Petroleum oil combustion fumes	DNA replication	KEGG/REACTOME	IGSEA/KGG	0.007/0.002
Hard alloys dust	Gap junction	KEGG	IGSEA/KGG	<0.001/0.01
	Calcium signaling pathway	KEGG	IGSEA/KGG	0.001/0.028
	Apoptosis	Wiki/Reactome	IGSEA/KGG	0.007/0.0009
Inorganic insulation dust	Rho signaling	GO:0035023/Reactome	INRICH/KGG	0.009/0.02
	Transcription	Reactome	KGG	0.00008
	MAPK pathway	KEGG	KGG	0.00008
	TCR pathway	PID	KGG	0.0001
Inorganic	GPCR pathway	Wiki	IGSEA	<0.001

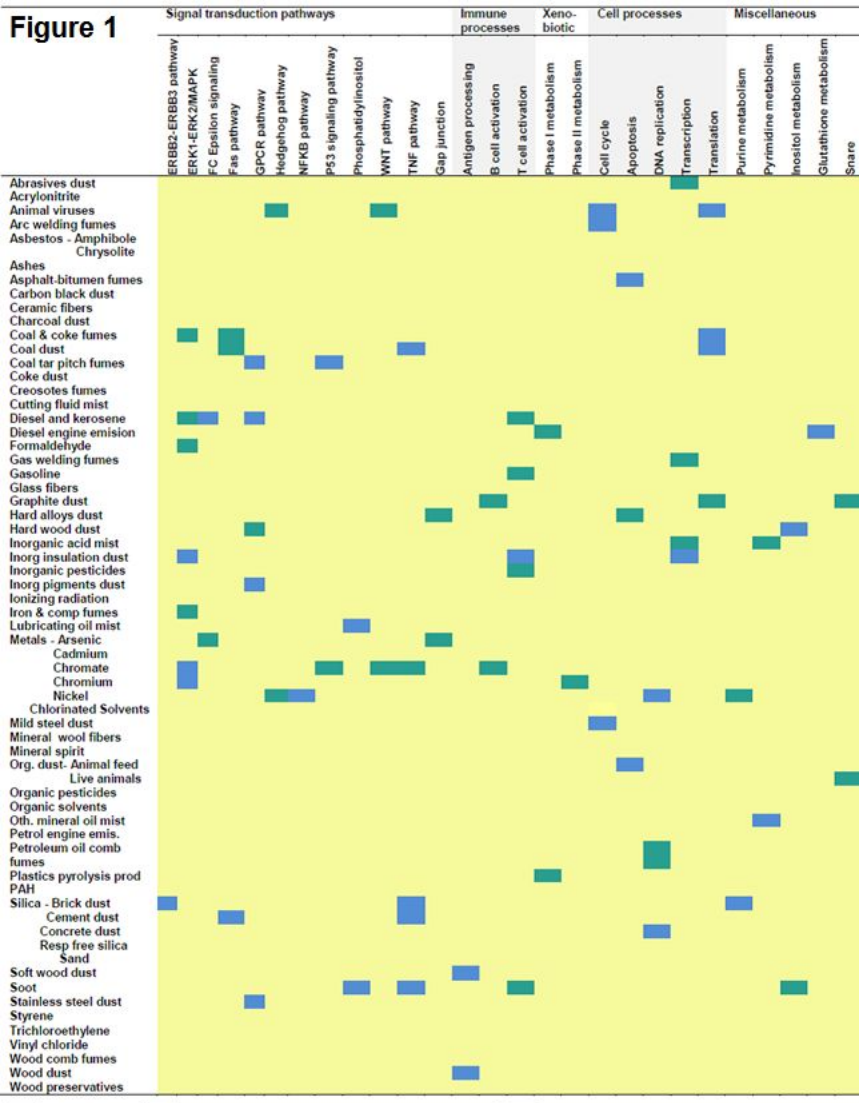
pigments dust				
Iron & compounds fumes	MAPK pathway	STKE/Reactome	IGSEA/KGG	0.019/0.02
	Arachidonic acid metabolism	KEGG	IGSEA/KGG	0.015/0.04
	Cell cycle	Wiki	IGSEA	0.001
Mild steel dust	Biosynthesis of steroids	KEGG	IGSEA	0.001
Stainless steel dust	GPCR pathway	Wiki	IGSEA	<0.001
	VEGF pathway	Biocarta	IGSEA	<0.001
Wood dust	Antigen processing and presentation	KEGG	IGSEA	0.008
Hard wood dust	GPCR pathway	Wiki/Biocarta	IGSEA/KGG	0.001/0.002
	Inositol metabolism	KEGG	IGSEA	<0.001
	PPARA pathway	Biocarta	KGG	0.00003
Soft wood dust	Antigen processing and presentation	KEGG	IGSEA	<0.001
Inorganic pesticide	NK cell mediated cytotoxicity	KEGG	INRICH	0.002
	T cell activation	GO:0031295/Biocarta	INRICH/IGSEA	0.001/<0.001
Formaldehyde	MAPK signaling pathway	Biocarta/KEGG	IGSEA/KGG	<0.001/0.05
Cutting fluid mist	IL2RB pathway	Biocarta	IGSEA/KGG	0.002/0.05
Inorganic acids mist	Pyrimidine metabolism	KEGG	IGSEA/KGG	0.01/0.01
	Transcription	GO:0030528/Reactome	INRICH/IGSEA/KGG	0.037/0.01/0.002
	Tight junction	KEGG/Reactome	INRICH/KGG	0.03/0.03
Lubricating oil mist	MTOR signaling pathway	KEGG	KGG	0.0002
	PI3K pathway	Reactome	KGG	0.00003
	Regulation of autophagy	KEGG	IGSEA	<0.001
Other mineral oil mist	Pyrimidine metabolism	KEGG	INRICH	0.001
Arc welding fumes	Cell cycle	Biocarta	IGSEA	<0.001
Gas welding fumes	Transcription	GO:0006350/KEGG	INRICH/KGG	0.02/0.0004
Animal viruses	Translation	KEGG	IGSEA	0.005
	Cell cycle	KEGG	IGSEA/KGG	0.001/0.0001
	Hedgehog signaling pathway	KEGG	IGSEA/KGG	0.016/0.001
	WNT signaling	SuperArray	IGSEA/KGG	0.003/0.003
Abrasives dust	Transcription	GO:0016481/KEGG	INRICH/IGSEA	0.005/0.002
	GTPase activity	GO:0003924	INRICH	0.007
	Adenylate cyclase pathway	Reactome	KGG	0.003
Plastics pyrolysis fumes	Oxidative phosphorylation	KEGG	IGSEA/KGG	0.005/0.03
	Xenobiotic metabolism by C450	KEGG	IGSEA/KGG	0.01/0.04

Abbreviations: IGSEA, Improved Gene Set Enrichment Analysis; KGG, Knowledge-based mining system for Genome-wide Genetic studies; KEGG, Kyoto Encyclopedia of Genes and Genomes; STKE, Signal Transduction Knowledge Environment; PID, Pathway Interaction Database

Figure Legends:

Figure 1. Pathway-based analysis using occupational exposure interactions

Figure 1



Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

Effect of Occupational Exposures on Lung Cancer Susceptibility: A Study of Gene-Environment Interaction Analysis

Jyoti Malhotra, Samantha Sartori, Paul Brennan, et al.

Cancer Epidemiol Biomarkers Prev Published OnlineFirst January 12, 2015.