A gene-based risk score for lung cancer susceptibility in smokers and ex-smokers

R P Young,1 R J Hopkins,1 B A Hay,1 M J Epton,2 G D Mills,3 P N Black,1 H D Gardner,1 R Sullivan,4 G D Gamble1

ABSTRACT

Background: Epidemiological and family studies suggest that lung cancer results from the combined effects of age, smoking and genetic factors. Chronic obstructive pulmonary disease (COPD) is also an independent risk factor for lung cancer and coexists in 40–60% of lung cancer cases.

Methods: In a two-stage case–control association study, genetic markers associated with either susceptibility or protection against lung cancer were identified. In a test cohort of 439 Caucasian smokers or ex-smokers, consisting of healthy smokers and lung cancer cases, 157 candidate single nucleotide polymorphisms (SNPs) were screened. From this, 30 SNPs were identified, the genotypes (codominant or recessive model) of which were associated with either the healthy smokers (protective) or lung cancer (susceptibility) phenotype. After genotyping of this 30-SNP panel in a second validation cohort of 491 subjects and using the same protective and susceptibility genotypes from our test cohort, a 20-SNP panel was selected on the basis of independent univariate analyses.

Results: Using multivariate logistic regression, including the 20 SNPs, it was also found that age, history of COPD, family history of lung cancer and gender were significantly and independently associated with lung cancer.

Conclusions: When numeric scores were assigned to both the SNP and demographic data, and sequentially combined by a simple algorithm in a risk model, the composite score was found to be linearly related to lung cancer risk with a bimodal distribution. Genetic data may therefore be combined with other risk variables from smokers or ex-smokers to identify individuals who are most susceptible to developing lung cancer.

Approximately 90% of people with lung cancer have a smoking history, yet only 10–15% of chronic smokers get lung cancer, suggesting that factors in addition to smoking exposure must be relevant.1 Epidemiological studies have identified age, smoking exposure, impaired lung function and family history as key risk factors for lung cancer.2 Genetic factors have been shown to play a modest role in determining susceptibility to lung cancer,3 4 most likely by conferring an inherent susceptibility (exaggerated or maladaptive response) to chronic inflammation from aero-pollutant exposure (most commonly smoking).5 6 Like many cancers, this provides the initial stimulus to tissue remodelling (eg, small airway disease and/or emphysema), DNA damage and impaired cell cycle control.7

Prospective studies have shown that ~20% of smokers develop chronic obstructive pulmonary disease (COPD), defined by spirometry (forced expiratory volume in 1 s (FEV1)/forced vital capacity <70% and/or FEV1<80% of predicted), whereas the majority have preserved lung function at, or close to, predicted values.8 Both prospective and retrospective studies show that spirometric evidence of COPD is found in 40–60% of smokers diagnosed with lung cancer.9 10 In contrast with smokers who maintain normal lung function, those with COPD have a 2–6-fold greater risk of lung cancer, independent of age and pack-years.9 11 These studies suggest that smokers with COPD are at an inherently increased risk of lung cancer (susceptible phenotype), whereas most smokers with normal lung function (estimated to be 80%) are at least risk (resistant phenotype).12 13

Genetic predisposition to lung cancer is likely to be both polygenic and heterogeneous, conferred by a variable combination of relatively common polymorphisms with low penetrance and modest effect sizes.12 13 Moreover, it is likely that important smoking-gene interactions underlie lung cancer,14 as seen in other smoking-related cancers (eg, bladder and stomach). As genetic variants associated with both COPD and lung cancer have been identified, and include the recently reported chromosome 15q25 gene locus,15 16 we suggest it is important to measure lung function in all participants of case–control studies of lung cancer. For both clinical and biostatistical reasons, screening the exposed controls will increase the power of the study to identify relevant genetic variants compared with studies in which the control group is unscreened.17

It is well known that non-genetic risk factors such as age, history of COPD and smoking history are very important and can be combined to develop risk-based tools for determining lung cancer susceptibility.10 19 Recently, genotype data from previously implicated prostate cancer susceptibility single-nucleotide polymorphisms (SNPs) were combined with family history to derive risk estimates for prostate cancer.20 The objective of this study was to use a similar approach to analysing data from a case–control study and show how genetic variants, previously showing small effects on lung cancer risk, can be combined in an algorithm with other known risk factors to derive a gene-based susceptibility score for lung cancer.

METHODS

Study population

This study was a two-stage case–control design conducted in three centres following the same recruitment protocol. Only people of Caucasian
ancestry were recruited (all four grandparents of Caucasian descent). Lung cancer cases were identified through hospital clinics between 2004 and 2007 using the following criteria: >40 years of age, minimum 15 pack-years of smoking, diagnosis confirmed on histological or cytological grounds, and limited to the following four histological subtypes: adenocarcinoma, squamous cell cancer, small cell cancer and non-small cell cancer (generally large cell or bronchoalveolar subtypes). The median time interval between diagnosis and recruitment was 3 months. Patients with lung cancer underwent blood sampling for DNA extraction, an investigator-administered questionnaire and spirometry with a portable spirometer (Easy-One; ndd Medizintechnik AG, Zurich, Switzerland) following American Thoracic Society (ATS) criteria. For patients with lung cancer who had already undergone surgery, results of preoperative lung function tests performed by the hospital laboratory (using ATS criteria) were sourced from the medical records.

Control subjects were recruited from the same communities as the cases using the following criteria: Caucasian ancestry (as defined above), age 45–80 years and past or current smoking history of a minimum of 15 pack-years. Controls were volunteers who met the above criteria and were identified through either a community mail-out or while attending a community-based club for older people located in the suburbs that were the referral base for the hospital clinics from which the lung cancer cases were recruited. All smoking controls underwent blood sampling, spirometry and the same investigator-administered questionnaire given to the cases. Controls with spirometry consistent with COPD were excluded (30% of those who volunteered). Informed written consent was obtained from both lung cancer patients and community-acquired healthy “smoker” controls. The study was approved by the local ethics committee. The questionnaire (modified from the ATS respiratory questionnaire) included data on demographic variables such as age, gender, medical history, family history of lung disease, active and passive tobacco exposure, and occupational aero-pollutant exposures.

**Selection and genotyping of SNPs**

After extensive review of both the lung cancer and COPD literature, polymorphisms with the following attributes were selected for initial screening in the test cohort: (a) SNPs in genes encoding proteins in pathways of cell-cycle control, oxidant response, apoptosis and airways inflammation; (b) SNPs known either to have functional effects on in vitro assays or to be either non-synonymous or in regulatory regions. In a test cohort of 459 smokers (run 1 recruited during 2003–2005: 259 lung cancer cases and 200 control smokers), 157 candidate SNPs were screened (available on request), and those where the difference in genotype frequencies exceeded a 20% magnitude difference and p value <0.20 were identified as part of our model-forming approach. Where the call rate (percentage of samples for which genotyping failed for technical reasons) fell below 95% for any cohort, the reading and/or genotyping of failed samples was repeated for that SNP; after retesting, SNPs with call rates <95% were not included in further analysis. SNPs were assigned as “protective” when the homozygote and/or heterozygote genotype for either allele was found more often in control smokers than lung cancer cases (in a recessive or codominant model). SNPs were assigned as “susceptible” when the homozygote and/or heterozygote genotype was found significantly more often in lung cancer cases than control smokers.

**Genotyping**

Genomic DNA was extracted from whole blood samples using standard salt-based methods. Purified genomic DNA was aliquoted (10 ng/µl concentration) into 96-well plates and genotyped on a Sequenom system (Sequenom Autoflex Mass Spectrometer and Samsung 24 pin nanodispenser) by the Australian Genome Research Facility (www.agrf.com.au) using sequences designed in-house (available on request) and recommended amplification and separation methods (iPLEX; www.sequenom.com).

Of the 157 candidate SNPs screened in our discovery cohort, 30 met the above criteria. These SNPs were genotyped in a second cohort of 491 smokers using identical recruitment methods (run 2 recruited during 2006–2007: 207 lung cancer cases and 284 control smokers). For all SNP assays, again a minimum call rate of 95% was required. This second validation cohort of lung cancer cases and control smokers was identical with the first groups with respect to demographic factors and lung cancer characteristics. On the basis of independent univariate analyses in run 1 and run 2 (consistency, direction and significance of association), a final panel of the 20 most discriminatory SNPs was selected (12 susceptibility SNPs and eight protective SNPs from the test panel of 30).

**Algorithm**

The assignment of a protective or susceptible SNP genotype was made from the test cohort data (run 1) and was strictly applied to the data from run 2. On the basis of an algorithm derived from our work on the genetics of COPD (unpublished data), a scoring system was applied to the genotypes for each of the susceptibility and protective SNPs. For each subject, a numerical value of −1 was assigned for each of the protective genotypes present among the protective SNPs and +1 for each of the susceptible genotypes present. Where an individual did not have either the protective or susceptibility genotype for that SNP, the score was 0 (ie, did not contribute to the genetic score). This approach is consistent with a recently published study in prostate cancer.20 Weighting the presence of specific susceptible or protective genotypes according to their individual odds ratios (ORs; from univariate regression) did not significantly improve the discriminatory performance of the raw SNP score (unpublished data).

**Lung cancer susceptibility score**

The approach of deriving an overall “susceptibility score” by combining independent risk factors is comparable to existing risk-scoring systems such as the Prostate Cancer Test, Framingham Score for coronary artery disease risk and the Gail Score for breast cancer.21–25 By using multivariate logistic and stepwise regression analysis, the 20-SNP panel was examined in combination with relevant non-genetic factors. This analysis of run 1 data identified age, family history of lung cancer and previous diagnosis of COPD as significant contributors to lung cancer susceptibility. In addition, and consistent with other case–control studies, female gender in our study was also associated with a small increased risk of lung cancer (p<0.01). However, we did not include gender in the final risk model, as its importance in prospective studies has been lacking.21 On the basis of a multivariate logistic regression analysis in run 1 (see results for combined analysis below), a score was assigned according to age, history of COPD and family history. These variables have been identified in other risk assessment tools for lung cancer susceptibility21 25 and improved
the discriminatory power of the SNP score data alone. As smoking exposure (pack-years) was a recruitment criterion for this study and comparable between cases and controls, it was not surprising to find that it made little contribution to this scoring system derived from our cohorts. The lung cancer susceptibility score was plotted with (a) the frequency of lung cancer and (b) the floating absolute risk (equivalent to OR) across the combined smoker/ex-smoker cohort.35-36

Statistical analysis
Patient characteristics in the cases and controls were compared by unpaired t tests for continuous variables and χ² test for discrete variables. Genotype and allele frequencies were checked for each SNP by Hardy–Weinberg equilibrium (tests that genotype frequencies were as expected from the allele frequencies). Population admixture was excluded by the population structure analysis on genotyping data from 40 unrelated SNPs.27 Distortions in the genotype frequencies were identified between cases and controls using 2 × 3 contingency tables. Genotype data (20-SNP panel) and the most relevant non-genetic variables were combined in a stepwise fashion to assess their combined effects on discriminating low and high risk (by OR and receiver operating characteristic (ROC) curve) by score quintile. The frequency distribution of the optimised lung cancer susceptibility score was compared across the cases and controls. Its clinical utility was assessed using ROC analysis, which assesses how well the model predicts risk across the score (ie, clinical performance of the score with respect to sensitivity, specificity and false positive rate). To assess the stability of the optimised risk model, a sensitivity analysis was performed in which age, gender and smoking dose were more stringently matched between cases and controls. The effect on sensitivity and performance of the lung cancer susceptibility score to the addition of non-genetic variables was also assessed by comparing ORs and ROC analyses.

RESULTS

Demographic variables and genotyping
Table 1 summarises the characteristics of the lung cancer cases and healthy control smokers. The 446 lung cancer cases from run 1 (n = 239) and run 2 (n = 207) were comparable (with respect to demographic characteristics, histology and staging) and similar to a large published series.28 Given the small difference in age, the 482 healthy control smokers (200 in run 1, 282 in run 2) were comparably exposed with respect to smoking and other aero-pollutants. The lower frequency of current smokers in the lung cancer group probably reflects coexisting COPD (higher quit rates), and longer duration of smoking in lung cancer cases reflects an older age. In a gene-by-smoking interaction model such as this, differences in smoking exposure are more likely to obscure effects (bias to the null) than generate effects. Consistent with the findings of others, the lung cancer cohort had higher rates of a family history of lung cancer (19% vs 9%) and history of COPD (29% vs 5%). The latter (5%) probably reflects a clinical diagnosis of COPD, based on symptoms but not spirometry, in smokers with asthma and/or chronic bronchitis. As expected, lung function was worse in the lung cancer cohort than the healthy smoker controls. Testing lung function in the lung cancer cases (performed within 3 months of diagnosis, in the absence of pleural effusions and before surgery) allows us to test for confounding by COPD (see below).

The observed genotypes for the 20 SNPs in this study were in Hardy–Weinberg equilibrium (table 2), thereby excluding significant genotyping error. The genotype frequencies for the controls were comparable to those from the International Hapmap Project (www.hapmap.org). The development of the lung cancer susceptibility score is described in the Methods section, and a summary of the 20-SNP panel univariate analysis is presented in table 3. Although six of the top 20 SNPs do not reach traditional levels of significance, they have been included in the panel because (a) in previous studies they have been shown to have functional effects, (b) they have been associated with COPD and/or lung cancer (see Discussion), (c) in combination they make a contribution to the performance of the susceptibility score, and (d) their inclusion recognises the likely genetic heterogeneity that exists in lung cancer case–controls studies. A SAS macro was used to estimate the false discovery rate (FDR) (Osborne JA, North Carolina State University; http://www2.sas.com/proceedings/sugi31/190-31.pdf) and produce a q statistic as the smallest p value that would be said to be statistically significant while preserving an overall 5% significance level.

Table 1 Summary of characteristics for the lung cancer and resistant smokers

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lung cancer patients (n = 446)</th>
<th>Healthy control smokers (n = 484)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>53%</td>
<td>60%</td>
<td>0.007</td>
</tr>
<tr>
<td>Age (years)</td>
<td>69 (10)</td>
<td>60 (10)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>167 (0.08)</td>
<td>170 (0.09)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>69 (15)</td>
<td>79 (15)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>History of COPD</td>
<td>29%</td>
<td>5%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-smokers (%)</td>
<td>7%</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Current smoking (%)</td>
<td>35%</td>
<td>48%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age started (years)</td>
<td>18 (4)</td>
<td>17 (3)</td>
<td>0.18</td>
</tr>
<tr>
<td>Years smoked</td>
<td>41 (12)</td>
<td>35 (11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pack-years</td>
<td>41 (25)</td>
<td>40 (19)</td>
<td>0.28</td>
</tr>
<tr>
<td>Cigarettes/day</td>
<td>20 (10)</td>
<td>24 (11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>History of other exposures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In utero smoke exposure</td>
<td>18%</td>
<td>17%</td>
<td>0.45</td>
</tr>
<tr>
<td>Mother smoked in childhood</td>
<td>37%</td>
<td>41%</td>
<td>0.03</td>
</tr>
<tr>
<td>Home ETS exposure as adult</td>
<td>79%</td>
<td>58%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Work ETS exposure</td>
<td>88%</td>
<td>63%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Work dust exposure</td>
<td>63%</td>
<td>47%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Work fume exposure</td>
<td>41%</td>
<td>38%</td>
<td>0.16</td>
</tr>
<tr>
<td>Asbestos exposure</td>
<td>23%</td>
<td>16%</td>
<td>0.02</td>
</tr>
<tr>
<td>Family history</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD</td>
<td>33%</td>
<td>28%</td>
<td>0.12</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>19%</td>
<td>9%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lung function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEV₁ (litres)</td>
<td>1.86 (0.48)</td>
<td>2.86 (0.68)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FEV₁ % predicted</td>
<td>73%</td>
<td>99%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FEV₁/FVC</td>
<td>64 (13)</td>
<td>78 (7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Spirometric COPD*</td>
<td>51%</td>
<td>0%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Values are frequency (given as %) or mean (SD).

* According to GOLD 2 criteria (FEV₁/FVC < 70% and FEV₁ % predicted < 80%).

COPD, chronic obstructive pulmonary disease; ETS, environmental tobacco smoke; FEV₁, forced expiratory volume in 1 s; FVC, forced vital capacity.
Such an approach is consistent with existing risk scores and places the SNP data in an appropriate clinical context. Gender described above (and its inclusion did not alter the AUC).

<table>
<thead>
<tr>
<th>SNP No</th>
<th>SNP name (rs)</th>
<th>Major</th>
<th>Minor</th>
<th>HWE observed genotypes (p Value)</th>
<th>Allele frequencies (study total = 930)</th>
<th>Study controls (n = 484)</th>
<th>Hapmap Caucasian*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A5 nACHR (rs 16989968)</td>
<td>0.65</td>
<td>0.35</td>
<td>0.57</td>
<td>0.69</td>
<td>0.31</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>CYP 2E1 (rs 2031920)</td>
<td>0.98</td>
<td>0.02</td>
<td>0.06</td>
<td>0.99</td>
<td>0.01</td>
<td>0.94</td>
</tr>
<tr>
<td>3</td>
<td>Interleukin-18 (rs 360721)</td>
<td>0.70</td>
<td>0.30</td>
<td>0.77</td>
<td>0.68</td>
<td>0.32</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>Interleukin-8 (rs 4073)</td>
<td>0.54</td>
<td>0.46</td>
<td>0.14</td>
<td>0.51</td>
<td>0.49</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>Interleukin 18 (rs 16944)</td>
<td>0.69</td>
<td>0.31</td>
<td>0.80</td>
<td>0.67</td>
<td>0.33</td>
<td>0.65</td>
</tr>
<tr>
<td>6</td>
<td>ITGA11 (rs 2306022)</td>
<td>0.90</td>
<td>0.10</td>
<td>0.07</td>
<td>0.91</td>
<td>0.09</td>
<td>0.93</td>
</tr>
<tr>
<td>7</td>
<td>NAT 2 (rs 179930)</td>
<td>0.71</td>
<td>0.29</td>
<td>0.87</td>
<td>0.69</td>
<td>0.31</td>
<td>0.71</td>
</tr>
<tr>
<td>8</td>
<td>α1-antichymotrypsin (rs 4934)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.54</td>
<td>0.47</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>9</td>
<td>Cerberus 1 (rs 10115703)</td>
<td>0.92</td>
<td>0.08</td>
<td>0.54</td>
<td>0.92</td>
<td>0.07</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>DAT1 (rs 6413429)</td>
<td>0.93</td>
<td>0.07</td>
<td>0.78</td>
<td>0.85</td>
<td>0.15</td>
<td>0.84</td>
</tr>
<tr>
<td>11</td>
<td>TNFR1 (rs 1139417)</td>
<td>0.57</td>
<td>0.43</td>
<td>0.96</td>
<td>0.56</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td>12</td>
<td>TLR9 (rs 5743836)</td>
<td>0.85</td>
<td>0.15</td>
<td>0.92</td>
<td>0.85</td>
<td>0.15</td>
<td>0.84</td>
</tr>
<tr>
<td>13</td>
<td>P73 (TP73) (rs 2273953)</td>
<td>0.75</td>
<td>0.25</td>
<td>0.93</td>
<td>0.78</td>
<td>0.22</td>
<td>0.85</td>
</tr>
<tr>
<td>14</td>
<td>SNAP3 (rs 1798985)</td>
<td>0.99</td>
<td>0.01</td>
<td>0.99</td>
<td>0.98</td>
<td>0.02</td>
<td>0.95</td>
</tr>
<tr>
<td>15</td>
<td>ITGB3 (rs 2317676)</td>
<td>0.93</td>
<td>0.07</td>
<td>0.51</td>
<td>0.91</td>
<td>0.09</td>
<td>0.90</td>
</tr>
<tr>
<td>16</td>
<td>DRD2 (rs 1799732)</td>
<td>0.90</td>
<td>0.10</td>
<td>0.67</td>
<td>0.89</td>
<td>0.11</td>
<td>0.90</td>
</tr>
<tr>
<td>17</td>
<td>BCL2 (rs 2279115)</td>
<td>0.51</td>
<td>0.49</td>
<td>0.60</td>
<td>0.53</td>
<td>0.47</td>
<td>0.57</td>
</tr>
<tr>
<td>18</td>
<td>XPD (ERCC2) (rs 13181)</td>
<td>0.61</td>
<td>0.39</td>
<td>0.90</td>
<td>0.59</td>
<td>0.41</td>
<td>0.67</td>
</tr>
<tr>
<td>19</td>
<td>REVI (REV1) (rs 3087386)</td>
<td>0.56</td>
<td>0.44</td>
<td>0.96</td>
<td>0.58</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>20</td>
<td>Fasl (TNFSF6) (rs 763110)</td>
<td>0.63</td>
<td>0.37</td>
<td>0.83</td>
<td>0.63</td>
<td>0.39</td>
<td>0.64</td>
</tr>
</tbody>
</table>

*Allele frequencies for Caucasians from www.hapmap.org. SNP, single nucleotide polymorphism.

Risk model development

In a multivariate logistic regression analysis that included the SNPs (individually), age (>60 years), family history of lung cancer (first-degree relative), gender and history of COPD, the OR for the susceptibility and protective SNPs was 1.1–3.2 and 0.20–0.80, respectively (the combined SNP score is independently related to lung cancer, p<0.001). The OR for age >60 years, family history of lung cancer and history of COPD was 3.5 (95% CI 2.5 to 4.9, p<0.001), 2.5 (95% CI 1.6 to 4.0, p<0.001) and 7.5 (95% CI 4.5 to 12.4, p<0.001), respectively (total area under the curve (AUC) = 0.80 where SNPs were included individually with adjustment for all variables). History of COPD in this model confers a high risk, in part due to differences in lung function derived from the study design. On the basis of these findings, and those from previously published studies,3–4, 9–11 we derived an optimised score by assigning scores to non-genetic variables as follows; (a) increasing age (age weighting equally applied to all), (b) where mean age, pack-years and gender were closely matched between cases and controls (referent group). The OR for age (referent group) and compares the lung cancer score with the referent group. The OR for SNPs alone (fig 2a), SNPs, family history and age (fig 2b) and SNPs, family history, age and COPD (fig 2c) spanned from 1 to 10 (p<0.001), 1 to 19 (p<0.001) and 1 to 28 (p<0.001), respectively, across the lung cancer scores when subjects were grouped approximately as heptiles or quintiles. Subgrouping by age band or histology did not alter this linear relationship between score and OR (data not shown). The lung cancer susceptibility score for lung cancer cases and controls shows a bimodal distribution on frequency distribution, indicating potential utility as screening test.29

Analysis of model sensitivity

To correct for the small differences in age, smoking status, COPD and gender mix between cases and controls, a subgroup (sensitivity) analysis was performed (a) limited to those >60 years of age (age weighting equally applied to all), (b) removing COPD from the model, and (c) where mean age, pack-years and gender were closely matched between cases and controls (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male respectively). A linear increase in OR across quintiles of the lung cancer susceptibility score (range 1–58, p<0.001) and 7.5 (95% CI 4.5 to 12.4, p<0.001), respectively. When just ''genetic factors'' are used in the risk model, the AUC or c statistic for run 1, run 2 and run 1 was 0.82, 0.75 and 0.79, respectively. The AUC in the total cohort spanned from 1 to 10 (p<0.001), 1 to 19 (p<0.001) and 1 to 28 (p<0.001), respectively, across the lung cancer scores when subjects were grouped approximately as heptiles or quintiles. Subgrouping by age band or histology did not alter this linear relationship between score and OR (data not shown). The lung cancer susceptibility score for lung cancer cases and controls shows a bimodal distribution on frequency distribution, indicating potential utility as screening test.29

ROC analysis

In a ROC analysis (n = 930) of the optimised model, we found that the AUC or c statistic for run 1, run 2 and run 1 was 0.82, 0.75 and 0.79, respectively. The AUC in the total cohort for the 20-SNP panel, age, family history of lung cancer and history of COPD on their own were 0.68, 0.70, 0.55 and 0.62, respectively. When just “genetic factors” are used in the risk model (SNPs + family history of lung cancer), as seen in the Prostate Cancer Study,20 the ORs span 1–10 across septiles and the AUC = 0.70 (with no contribution from age and COPD).
On stepwise analysis, age and the SNP panel make the greatest contribution to the AUC (SNPs = 0.68, age + SNPs = 0.76, age + SNPs + family history = 0.77), with history of COPD making a small additional contribution (total AUC = 0.79) (fig 3). Using an FDR analysis, 12 SNPs were identified as being significantly associated with lung cancer and, when combined with age and family history, derived an AUC = 0.75. When gender was included in the model, the AUC was not improved.

Inclusion of COPD

As smokers with normal lung function were selected as controls (resistant or lowest risk phenotype), and history of COPD was included in the model, it is necessary to examine the effect of including COPD in the model. In the ROC analysis, history of COPD alone was a modest discriminator (AUC = 0.62) and added little to the other variables in the combined model (increased AUC from 0.77 to 0.79). When history of COPD was removed from the model, the ORs span 1–19 across quintiles (p<0.001) and performance characteristics are minimally affected (AUC = 0.77). When history of COPD was removed from the model, the ORs span 1–19 across quintiles (p<0.001). Most importantly, the model was assessed by comparing the smoking controls and lung cancer cases subgrouped according to lung function (fig 4a,b). This shows that (a) the distribution of the susceptibility score was comparable between lung cancer cases divided by those with high or low lung function (fig 4a) and (b) when people with COPD (based on spirometry) are removed from the analysis (leaving smoking controls compared with lung cancer cases with normal lung function), the bimodal distribution is not affected (fig 4b). We conclude that the risk model is not significantly affected after adjustment for differences in COPD prevalence between cases and controls.

### Table 3  Genotypes and results of univariate analysis

<table>
<thead>
<tr>
<th>SNP (rs No)*</th>
<th>Genotype</th>
<th>Lung cancer</th>
<th>Smoking controls</th>
<th>Call rate (%)</th>
<th>Univariate OR (95% CI)</th>
<th>p Value</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Acetyl cysteine transferase 2 (rs 1799830)</td>
<td>AA</td>
<td>239 (56%)</td>
<td>222 (47%)</td>
<td>97</td>
<td>1.4 (1.1 to 1.9)</td>
<td>0.006</td>
<td>Susceptibility</td>
</tr>
<tr>
<td>α5-nAChR (rs 16969968)</td>
<td>AA</td>
<td>68 (16%)</td>
<td>45 (9%)</td>
<td>1.8 (1.2 to 2.7)</td>
<td>0.004</td>
<td>Susceptibility</td>
<td></td>
</tr>
<tr>
<td>CYP 2E1 (rs 2031920)</td>
<td>AG/GG</td>
<td>361 (84%)</td>
<td>426 (91%)</td>
<td>95</td>
<td>2.1 (1.0 to 4.3)</td>
<td>0.03</td>
<td>Susceptibility</td>
</tr>
<tr>
<td>Interleukin-18 (rs 360721)</td>
<td>AA/AG</td>
<td>224 (51%)</td>
<td>269 (56%)</td>
<td>96</td>
<td>1.4 (1.1 to 1.9)</td>
<td>0.009</td>
<td>Susceptibility</td>
</tr>
<tr>
<td>Interleukin-8 (rs 16944)</td>
<td>AA</td>
<td>14 (3%)</td>
<td>6 (1%)</td>
<td>98</td>
<td>2.6 (0.9 to 7.6)</td>
<td>0.04</td>
<td>Susceptibility</td>
</tr>
<tr>
<td>ITGA11 (rs 2308022)</td>
<td>AA</td>
<td>14 (3%)</td>
<td>6 (1%)</td>
<td>98</td>
<td>2.0 (0.7 to 5.9)</td>
<td>0.14</td>
<td>Susceptibility</td>
</tr>
<tr>
<td>N-Acetyl cysteine transferase 2 (rs 1799830)</td>
<td>AG/GG</td>
<td>361 (84%)</td>
<td>426 (91%)</td>
<td>95</td>
<td>2.1 (1.0 to 4.3)</td>
<td>0.03</td>
<td>Susceptibility</td>
</tr>
</tbody>
</table>

*OMIM nomenclature.
SNP, single nucleotide polymorphism.

On stepwise analysis, age and the SNP panel make the greatest contribution to the AUC (SNPs = 0.68, age + SNPs = 0.76, age + SNPs + family history = 0.77), with history of COPD making a small additional contribution (total AUC = 0.79) (fig 3). Using an FDR analysis, 12 SNPs were identified as being significantly associated with lung cancer and, when combined with age and family history, derived an AUC = 0.75. When gender was included in the model, the AUC was not improved.

Inclusion of COPD

As smokers with normal lung function were selected as controls (resistant or lowest risk phenotype), and history of COPD was included in the model, it is necessary to examine the effect of including COPD in the model. In the ROC analysis, history of COPD alone was a modest discriminator (AUC = 0.62) and added little to the other variables in the combined model (increased AUC from 0.77 to 0.79). When history of COPD was removed from the model, the ORs span 1–19 across quintiles (p<0.001) and performance characteristics are minimally affected (AUC = 0.77). The model was also tested in young smokers (< 60 years old), in whom COPD prevalence was only 3% and the ORs spanned 1–16 (p<0.001). Most importantly, the model was assessed by comparing the smoking controls and lung cancer cases subgrouped according to lung function (fig 4a,b). This shows that (a) the distribution of the susceptibility score was comparable between lung cancer cases divided by those with high or low lung function (fig 4a) and (b) when people with COPD (based on spirometry) are removed from the analysis (leaving smoking controls compared with lung cancer cases with normal lung function), the bimodal distribution is not affected (fig 4b). We conclude that the risk model is not significantly affected after adjustment for differences in COPD prevalence between cases and controls.
DISCUSSION

This study has used a two-stage case–control candidate gene approach and identified a panel of protective and susceptibility SNPs that individually confer only small effects (OR ranging from 0.3 to 2.6). This is very much in keeping with the experience from case–control association studies to date. Consistent with existing risk models, relevant factors were combined using an algorithm (in this study including SNP data) to derive a susceptibility score on a simple linear scale. This study design, and the algorithmic approach that underlies our lung cancer susceptibility score, takes into account important epidemiological observations relevant to genetic predisposition to lung cancer. Firstly, that, although smoking exposure is for the majority a prerequisite to developing lung cancer, increasing age, smoking dose and poor lung function have important independent effects on lung cancer susceptibility. Secondly, the genetic factors underlying lung cancer risk are likely to be both polygenic and heterogeneous, conferred by a variable combination of genetic variants (ie, SNPs with low penetrance and small effect sizes). Thirdly, genetic factors may confer either a protective or susceptibility phenotype to lung cancer. Here we report a 20-SNP panel which, combined with family history, define risk (OR) across quintiles ranging from 1 to 10 with an AUC of 0.70. A risk tool with greater clinical utility can be derived by including age and presence of COPD to identify those at greatest susceptibility to lung cancer (OR range 1–28 and AUC = 0.79).

Several other important factors relevant to the genetic epidemiology of lung cancer have been considered in the design of this study. We sought to minimise false-positive results in a number of ways. The most important of these was to internally validate our findings using a two-stage design with an initial test cohort (run 1) to identify SNPs of potential interest. We then tested only those SNPs in a second cohort of cases and controls (run 2) using univariate and multivariate analysis to rank the SNPs under both conditions. Secondly, population stratification was excluded, and, thirdly, the presence of genotyping error was minimised through Hardy–Weinberg equilibrium analysis (see Methods) and by the exclusion of SNPs with <95% call rate (fails on genotyping are invariably genotype specific, thus generating false-positive associations). With respect to important confounding factors, our lung cancer cases and healthy smoking controls were matched for smoking exposure (pack-years). They were also similar with respect to gender and age mix. When the combined cohort was subgrouped by age band, the lung cancer susceptibility score maintained its discriminating utility across all groups. It was concluded that lung function was not confounding the results of this study, as the distribution of the lung cancer susceptibility score across the lung cancer cases, subdivided by normal and low lung function, showed no significant difference. The same could not be said of previously published case–control studies in which lung function was not measured.

However, weaknesses in this study include the modest size of the cohorts, borderline significance of some SNPs in the absence.
of correction, cross-sectional design, and recruitment limited to Caucasians. Moreover, it is accepted that, by selecting a control population with normal lung function (but comparable exposure) and including COPD (history) in the score, we will increase the score in those with lung cancer compared with controls (5% vs 29%). However, although this increases the magnitude of the difference between cases and controls (reflected in the ORs), it contributes little to the performance of the score (adds 0.02 to the AUC; see the Results section). Moreover, when subjects with COPD are excluded from the analysis (fig 4b, smoking controls versus lung cancer cases with normal lung function), the discriminating utility of the score is unaffected. In addition, in the youngest age band (confined to cases and controls ≤60 years old), the prevalence of COPD (history) was only 3% (little effect from COPD weighting), there was no age weighting, and the susceptibility score was still a good discriminator (OR spans 1–16, p<0.001). We argue that screening individuals with COPD (based on spirometry) out of the controls has the following advantages: (1) best reflects the majority of smokers with no COPD estimated at 80%; (2) best reflects the majority of smokers who will not develop lung cancer (resistant phenotype) estimated to be 80–90% (thereby minimising the dilutional effects of including patients with COPD, ie, misclassification); (3) best suited to identifying “protective” SNPs by comparing exposed individuals at either end of the risk spectrum. That said, replication

Figure 3  Distribution of the lung cancer susceptibility (risk) score between cases and controls. (a) Single-nucleotide polymorphisms (SNPs) only, (b) SNPs, family history and age, and (c) SNPs, family history, age and history of chronic obstructive pulmonary disease. Ca, cancer.
Figure 4  (a) Frequency distribution of the lung cancer score among controls and lung cancer cases divided according to low and normal lung function. (b) Frequency distribution of lung cancer score among controls and lung cancer cases with normal lung function. Ca, cancer; FEV₁, forced expiratory volume in 1s.

Using an unselected control group (in which COPD prevalence would be 10% or more) might better reflect an unselected at-risk population and, as expected, reduce differences in the susceptibility score between cases and controls (dilutional effect). Although population stratification was formally tested, and our population is confined to Caucasians (where population admixture is less of an issue), it is possible that this remains a problem. A further limitation of the study is that, although the cases and controls were arguably representative, not all variables were precisely matched (eg, age, gender and smoking patterns). We reanalysed our data in a closely matched cohort (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs controls, respectively) and found the performance of the susceptibility score across quintiles was unchanged (OR range 0.68–0.85). Twelve of these SNPs have been associated with lung cancer in other cohorts. It is likely that other SNPs from as yet unidentified genes will be identified in the future. To assess further the utility of the lung cancer susceptibility score, a prospective study is in progress. To date, the lung cancer cases (n = 43) have the same mean and distribution as the lung cancer cases reported in this study (unpublished data). Further case–control and functional studies will be needed to further explore the role of these SNPs in lung cancer susceptibility.

We propose that clinical utility of genotype data requires that many SNPs are analysed and their effects combined with other epidemiological factors of relevance. Using an unselected control group (in which COPD prevalence would be 10% or more) might better reflect an unselected at-risk population and, as expected, reduce differences in the susceptibility score between cases and controls (dilutional effect). Although population stratification was formally tested, and our population is confined to Caucasians (where population admixture is less of an issue), it is possible that this remains a problem. A further limitation of the study is that, although the cases and controls were arguably representative, not all variables were precisely matched (eg, age, gender and smoking patterns). We reanalysed our data in a closely matched cohort (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs controls, respectively) and found the performance of the susceptibility score across quintiles was unchanged (OR range 0.68–0.85). Twelve of these SNPs have been associated with lung cancer in other cohorts. It is likely that other SNPs from as yet unidentified genes will be identified in the future. To assess further the utility of the lung cancer susceptibility score, a prospective study is in progress. To date, the lung cancer cases (n = 43) have the same mean and distribution as the lung cancer cases reported in this study (unpublished data). Further case–control and functional studies will be needed to further explore the role of these SNPs in lung cancer susceptibility.

We propose that clinical utility of genotype data requires that many SNPs are analysed and their effects combined with other epidemiological factors of relevance. Using an unselected control group (in which COPD prevalence would be 10% or more) might better reflect an unselected at-risk population and, as expected, reduce differences in the susceptibility score between cases and controls (dilutional effect). Although population stratification was formally tested, and our population is confined to Caucasians (where population admixture is less of an issue), it is possible that this remains a problem. A further limitation of the study is that, although the cases and controls were arguably representative, not all variables were precisely matched (eg, age, gender and smoking patterns). We reanalysed our data in a closely matched cohort (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs controls, respectively) and found the performance of the susceptibility score across quintiles was unchanged (OR range 0.68–0.85). Twelve of these SNPs have been associated with lung cancer in other cohorts. It is likely that other SNPs from as yet unidentified genes will be identified in the future. To assess further the utility of the lung cancer susceptibility score, a prospective study is in progress. To date, the lung cancer cases (n = 43) have the same mean and distribution as the lung cancer cases reported in this study (unpublished data). Further case–control and functional studies will be needed to further explore the role of these SNPs in lung cancer susceptibility.

We propose that clinical utility of genotype data requires that many SNPs are analysed and their effects combined with other epidemiological factors of relevance. Using an unselected control group (in which COPD prevalence would be 10% or more) might better reflect an unselected at-risk population and, as expected, reduce differences in the susceptibility score between cases and controls (dilutional effect). Although population stratification was formally tested, and our population is confined to Caucasians (where population admixture is less of an issue), it is possible that this remains a problem. A further limitation of the study is that, although the cases and controls were arguably representative, not all variables were precisely matched (eg, age, gender and smoking patterns). We reanalysed our data in a closely matched cohort (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs controls, respectively) and found the performance of the susceptibility score across quintiles was unchanged (OR range 0.68–0.85). Twelve of these SNPs have been associated with lung cancer in other cohorts. It is likely that other SNPs from as yet unidentified genes will be identified in the future. To assess further the utility of the lung cancer susceptibility score, a prospective study is in progress. To date, the lung cancer cases (n = 43) have the same mean and distribution as the lung cancer cases reported in this study (unpublished data). Further case–control and functional studies will be needed to further explore the role of these SNPs in lung cancer susceptibility.

We propose that clinical utility of genotype data requires that many SNPs are analysed and their effects combined with other epidemiological factors of relevance. Using an unselected control group (in which COPD prevalence would be 10% or more) might better reflect an unselected at-risk population and, as expected, reduce differences in the susceptibility score between cases and controls (dilutional effect). Although population stratification was formally tested, and our population is confined to Caucasians (where population admixture is less of an issue), it is possible that this remains a problem. A further limitation of the study is that, although the cases and controls were arguably representative, not all variables were precisely matched (eg, age, gender and smoking patterns). We reanalysed our data in a closely matched cohort (n = 450: 72 vs 69 years, 45 vs 43 pack-years and 70% vs 70% male for cases vs controls, respectively) and found the performance of the susceptibility score across quintiles was unchanged (OR range 0.68–0.85). Twelve of these SNPs have been associated with lung cancer in other cohorts. It is likely that other SNPs from as yet unidentified genes will be identified in the future. To assess further the utility of the lung cancer susceptibility score, a prospective study is in progress. To date, the lung cancer cases (n = 43) have the same mean and distribution as the lung cancer cases reported in this study (unpublished data). Further case–control and functional studies will be needed to further explore the role of these SNPs in lung cancer susceptibility.
Main messages

- Lung cancer results from the combined effects of smoking and genetic susceptibility.
- Chronic obstructive pulmonary disease is a common pre-existing and independent risk factor for lung cancer.
- Genetic susceptibility for lung cancer includes genetic variants (single nucleotide polymorphisms (SNPs)) conferring reduced risk (“protective”) best identified using a healthy smoking cohort.
- Genetic susceptibility for lung cancer results from the combined effects of genetic variants (SNPs) conferring either susceptibility or protective predisposition.
- Genetic and non-genetic variables can be combined to give a global risk score for susceptibility to lung cancer.

Current research questions

- Can the lung cancer susceptibility score be validated in smokers and ex-smokers in prospective studies and other populations?
- Will use of the lung cancer susceptibility score improve patient outcomes to reduce risk of lung cancer and/or detect lung cancer at a treatable stage?

ACKNOWLEDGEMENTS

References


Use the “My folders” feature to save and organise articles you want to return to quickly—saving space on your hard drive. You can also save searches, which will save you time. You will only need to register once for this service, which can be used for this journal or all BMJ Journals, including the BMJ.

Save your favourite articles and useful searches
A gene-based risk score for lung cancer susceptibility in smokers and ex-smokers


Postgrad Med J 2009 85: 515-524
doi: 10.1136/pgmj.2008.077107

Updated information and services can be found at:
http://pmj.bmj.com/content/85/1008/515

These include:

References
This article cites 66 articles, 26 of which you can access for free at:
http://pmj.bmj.com/content/85/1008/515#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

Editor's choice (139)
Lung cancer (oncology) (17)
Lung cancer (respiratory medicine) (17)
Epidemiology (401)
Genetics (135)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/