Cryptic Haplotypes of SERPINA1 Confer Susceptibility to Chronic Obstructive Pulmonary Disease

Sally Chappell, Leslie Daly, Kevin Morgan, Josef Roca, Roberto Rabinovich, Ann Millar, Seamas C. Donnelly, Vera Keatings, William MacNee, Jan Stolk, Pieter Hiemstra, Massimo Miniati, Simonetta Monti, Clare M. O’Connor, and Noor Kalsheker

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality and morbidity worldwide. While cigarette smoking is a major cause of COPD, only 15% of smokers develop the disease, indicating major genetic influences. The most widely recognized candidate gene in COPD is SERPINA1, although it has been suggested that SERPINA3 may also play a role. To detect cryptic genetic variants that might contribute to disease, we identified 15 SNP haplotype tags from high-density SNP maps of the two genes and evaluated these SNPs in the largest case-control genetic study of COPD conducted so far. For SERPINA1, six newly identified haplotypes with a common backbone of five SNPs were found to increase the risk of disease by six- to 50-fold, the highest risk of COPD reported to date. In contrast, no haplotype associations for SERPINA3 were identified. Hum Mutat 27(1), 103–109, 2006.

KEY WORDS: α1-antitrypsin; SERPINA1; α1-antichymotrypsin; SERPINA3; COPD; SNP; polymorphism; haplotype

The Supplementary Material referred to in this article can be accessed at www.interscience.wiley.com/pages/1059-7794/suppimat. Received 9 June 2005; accepted revised manuscript 6 September 2005.

*Correspondence to: Professor Noor Kalsheker, Division of Clinical Chemistry, University Hospital, Nottingham, NG7 2UH, United Kingdom. E-mail: noor.kalsheker@nottingham.ac.uk

Grant sponsor: European Union 5th Framework Programme; Grant number: QLGI-CT-2001-01012.

Clare M. O’Connor and Noor Kalsheker are joint senior authors.

DOI 10.1002/humu.20275

Published online 8 November 2005 in Wiley InterScience (www.interscience.wiley.com).
[Reeves et al., 2002]. If sufficient inhibitor is not present, uninhibited enzyme activity can destroy the connective tissue of the lung, resulting in loss of normal tissue architecture and function. Given the role of this gene in COPD, there is a high prior probability of cryptic variation in the SERPINA1 gene contributing to this disease. Genetic studies of SERPINA1 variation other than the Z variant have yielded controversial results, which have been attributed to small sample sizes and a lack of adequate coverage of the gene [Poller et al., 1990; Sandford et al., 1997]. Similarly, a number of small case-control studies reported inconsistent results on the role of the other major circulating antiprotease, SERPINA3 (more commonly known as α-1-antichymotrypsin), in COPD [Poller et al., 1992; Sandford et al., 1998].

In the largest case-control study reported to date, we screened 1,018 COPD patients and 911 nondiseased control smokers recruited at six European centers for nine SNPs that define the common haplotypes for the SERPINA1 gene, plus the Pi S variant, which is the most common known mutation that causes mild deficiency in Europeans, and five SNPs that define common haplotypes of the SERPINA3 gene.

**MATERIALS AND METHODS**

**Subjects**

The numbers of cases and controls recruited from each center were as follows: Barcelona, 70 controls and 138 cases; Bristol, 151 controls and 129 cases; Dublin, 195 controls and 196 cases; Edinburgh, 81 controls and 169 cases; Leiden, 216 controls and 188 cases; and Pisa, 198 controls and 198 cases. Approval for the study was obtained from the appropriate committees at each recruitment center (el Comite de Investigacion del Hospital Clinic, Barcelona; Southmead Local Research Ethics Committee, Bristol; St. Vincent's Hospital Ethics and Medical Research Committee, Dublin; Lothian Research Ethics Committee, Edinburgh; de Commissie Medische Ethiek van het LUMC, Leiden; and il Comitato Etico, Azienda Ospedaliera Pisana, Pisa). Informed consent was obtained from all subjects.

The criteria for patient recruitment were a firm clinical diagnosis of stable COPD, airflow limitation as indicated by FEV1 < 70% normal predicted values (calculated as recommended [Crapo et al., 1981] and FEV1/FVC < 70%, no significant reversibility on bronchodilation [Brand et al., 1992], and a smoking history of ≥20 pack-years. Patients were excluded from the study if they had an established diagnosis of asthma, established obstructive syndrome or lung cancer, a history of smoking with center were included. This approach gave us an increase in power by accounting for the matched recruitment, we adjusted for any residual confounding due to smoking. Age was categorized into four groups (45-54 years; 55–64 years; 65–69 years; 70+ years), and smoking was categorized as heavy or light based on a definition of heavy smoking that was age, sex, and smoking with center were included. This approach was also used to correct haplotype-disease associations (see below).

**Analysis of Genetic Variation in Population Groups**

Each of the SNPs in SERPINA1 and SERPINA3 haplotype-tagging SNPs was also carried out at the MRC Geneservice (Babraham, UK) using PCR with fluorescently-labeled Taqman probes (Vic or Fam labels). Genotyping of the study population for the SERPINA1 and SERPINA3 haplotype-tagging SNPs was carried out at the MRC Geneservice (Babraham, UK) using PCR with fluorescently-labeled Taqman probes (Vic or Fam labels). The primer and probe sequences are available on request. As a quality control measure, 44 samples of known genotype determined by sequencing were included. These genotypes were unknown to the MRC but were known at the source. When discrepancies were noted the analysis was repeated. This resulted in 100% concordance for all the assays.

The SNPs are numbered with respect to the GenBank genomic DNA sequences AL132708.3 for SERPINA1, and AL049839.3 for SERPINA3.

**Identifying Haplotype-Tagging SNPs and Genotyping the Study Population**

We previously identified nine haplotype-tagging SNPs that define common haplotypes for SERPINA1 via high-density SNP mapping [Chappell et al., 2004]. In the current study we also screened for the Pi S variant (SNP 7) because even though it does not occur at a frequency of > 5%, it is the most common known deficiency variant in Europeans. Haplotype-tagging SNPs for the SERPINA3 gene were identified using a strategy similar to that used for SERPINA1. Briefly, SNPs were identified and confirmed by bidirectional sequencing of all exons and 5’ flanking regions in 44 samples. SNPs that occurred at more than 5% were genotyped in 291 samples by the MRC Geneservice (Babraham, UK) using PCR with fluorescently-labeled Taqman probes (Vic or Fam labels).
Haplotyping Analysis

For a full analysis of the relationships between COPD and the SERPINA1 and -A3 SNPs, we used a staged, haplotype-based approach. It was first necessary to identify which SNP groupings within each gene should be subjected to the haplotype analysis. There are a number of approaches for detecting haplotype groupings in a gene, but all are based on defining a group as a block of contiguous SNPs [Zhang and Jin, 2003]. The objective of these methods is to identify haplotype-tagging SNPs that can be used to reduce the number of SNPs to genotype while still achieving a large percentage coverage of variation within the gene. We were interested in identifying haplotypes that differed in frequency between COPD cases and controls, using haplotype-tagging SNPs already identified from high-density mapping of the SERPINA1 and -A3 genes. Since there is no reason to suspect that disease-associated haplotypes might be combinations of contiguous SNPs, and there are no proven shortcut methods for identifying such groupings, we screened all possible SNP groups in each gene for differences between cases and controls. This was done by performing an omnibus test [Zhao et al., 2000] as implemented in SAS/Genetics PROC HAPLOTYPE. Exact P-values based on a permutation test were employed. We used a screening value of 1,000 permutations to initially choose SNP groups with exact P-values $\leq 0.017$. This ensured that (based on 0.0099 being the exact 95% lower binomial confidence limit for 17 out of 1,000) a nominal P-value of $< 0.01$ was achieved. The P-value was then recalculated for these groups with 10,000 permutations, and the final groups were chosen based on a P-value of $< 0.01$.

Once SNP groups that were potentially related to COPD were identified, employing the standard EM algorithm, PROC HAPLOTYPE was used to generate all possible haplotype patterns within these groups and to compare the distributions between cases and controls. Exact P-values based on 10,000 permutations were generated.

Haplotypes with a frequency of over 0.5% in cases and controls combined that differed between cases and controls at the $p < 0.001$ level were considered relevant. This ensured that extremely rare haplotypes were excluded from consideration.

Confounder Adjustment

Because SNP data are unphased, it is not possible to assign haplotypes to individuals. PROC HAPLOTYPE uses the expectation maximization (EM) algorithm to determine the frequency of each haplotype in cases and controls, and performs an overall significance test [Zhao et al., 2000]. Without haplotype data at an individual level, however, only a limited confounder adjustment can be performed using a stratified approach. A recently described technique, termed haplotype trend regression (HTR), utilizes regression methods for confounder adjustment by assigning the probability of having a particular haplotype to each individual and substituting that probability for the usual 1/0 coded binary predictor variable of haplotype present or absent [Zaykin et al., 2002]. However, applying this technique to our data yielded odds ratios that were severely biased upwards, so we used an alternative unbiased approach. This approach uses a weighted logistic regression with the probability of having a particular haplotype defining the weights (see Supplementary Appendix S1; available online at www.interscience.wiley.com/pages/1059-7794 suppmat). The haplotype probabilities were obtained using PROC HAPLOTYPE and used to adjust haplotype odds ratios and significance levels for the effects of center, age, sex, and smoking. Odds ratios (adjusted and nonadjusted) were based on the logistic regression and are the odds of disease in those with a particular haplotype relative to the odds in those without that haplotype.

Assessment of Population Stratification

To assess the effects of stratification we used data obtained from screening the same study population for 20 SNPs in four unrelated genes: y-glutamyl cysteine synthetase catalytic and regulatory subunits, tumor necrosis factor, and epoxide hydrolase. Allelic association was first tested for each of the 20 SNPs using 2 $\times$ 2 contingency tables. The individual chi-square test statistics were summed for the 20 SNPs to give an overall summary chi-square statistic [Pritchard and Rosenberg, 1999].

Multiple Testing

Given the role of the SERPINA1 gene in COPD, the prior probability of cryptic genetic variation contributing to disease would be considered to be high, and as a result the issue of multiple testing is not as problematic as it is in genomewide scans, where no a priori assumptions are made. We did not explicitly account for multiple testing in the calculation of P-values in this study; rather, we chose to use a stringent P-value and provide details of the analysis.

RESULTS

Identification of Haploype-Tagging SNPs for SERPINA3

High-density SNP mapping of the SERPINA3 gene identified five SNPs that occurred with a frequency of $> 5%$. These SNPs produce six major haplotypes representing almost 100% coverage of the variation observed in the gene (Table 1).

Allele and Genotype Frequencies in COPD and Control Populations

The locations of the 10 SERPINA1 and five SERPINA3 SNPs screened for in the study population are indicated in Figure 1. The allele frequencies for the minor variant of each of the 10 SERPINA1 and five SERPINA3 SNPs in the control and COPD groups are shown in Table 2. For ease of presentation, the SNPs are numbered sequentially relative to the coordinates of the human genomic DNA sequences AL132708.3 for SERPINA1 and AL049839.3 for SERPINA3. The position and base change of each SNP relative to the hepatocyte transcription start site are also indicated in parentheses. SNP g.135728A $\rightarrow$ T (SNP 7) of SERPINA1 is the Pi S variant.

The initial comparison of allele frequencies indicated that SERPINA1 SNPs 1 and 7 differed significantly between controls and COPD groups. However, when corrected for smoking, age, gender, and center, only SNP 1 retained borderline significance (Table 2). There were no differences between COPD and control groups in the frequency of any of the five SERPINA3 SNPs before or after confounder adjustment (Table 2).

Only minor differences were detected in the frequencies of SNPs across the six recruitment sites for all 15 SNPs, with the exception of SERPINA1 SNP 10, which varied from 4.85% in Edinburgh to 14.18% in Pisa. Previously published data showed population differences for this SNP, with a higher frequency in eastern Europe [Samichuk et al., 1997]. There were, however, no significant differences in the frequency of this particular SNP.
between the cases (14.18%) and controls (13.26%) in the Pisa
group or for any other population group. All of the SNPs reported
were in HWE in both cases and controls.

The linkage disequilibrium (LD) results for
SERPINA1 in the
control group are shown in Table 3, confirming in this large
population the low values of LD that were previously noted in a
smaller group for SERPINA1 [Chappell et al., 2004]. The strength
of LD between SNPs in SERPINA3 is high in comparison to
SERPINA1 (Table 3).

Haplotype Analysis

For SERPINA1 the initial analysis identified seven SNP groups
that differed significantly between COPD cases and controls

<table>
<thead>
<tr>
<th>Table 1. Haplotypes of the SERPINA3 Gene in Normal Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>g.44695G &gt; T (-12799)</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>G</td>
</tr>
</tbody>
</table>

Rare Haplotypes (<1%)

Numbering is with respect to the GenBank sequence AL049839.3, with numbering with respect to the hepatocyte transcription start site shown in parentheses.

![Diagram of SERPINA1 and SERPINA3 genes with SNPs and haplotypes](image)

**FIGURE 1.** Location of SNPs in SERPINA1 and -A3. Shaded boxes represent coding exons (not to scale).

<table>
<thead>
<tr>
<th>Table 2. Allele Frequency of SERPINA1 and SERPINA3 SNP Variants in Controls and COPD Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP No</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td><strong>SERPINA1</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td><strong>SERPINA3</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

aCrude P-values.
bP-values corrected by logistic regression for age, gender, center, and smoking.

Numbering is with respect to the GenBank sequence AL132708.3 for SERPINA1 and AL049839.3 for SERPINA3, with numbering with respect to the hepatocyte transcription start site shown in parentheses.
There were no significant (P < 0.01) SNP groups of sizes 10, nine, four, three, or two. It was of note that none of the groups consisted of contiguous SNPs only.

The second stage of analysis for SERPINA1 identified six haplotypes that were risk factors for COPD. These consisted of one 5-SNP, two 6-SNP, two 7-SNP, and one 8-SNP haplotypes (Table 5). It was of note that regardless of haplotype size, the SNP variant present at each SNP site was the same.

For SERPINA3 the omnibus analysis indicated that no SNP group differed significantly (P < 0.01) between COPD cases and controls. As recommended by SAS Institute Inc. (Cary, NC), no further haplotype analysis was carried out on this gene.

Population Stratification

The overall chi-squared value for the 20 SNPs in unrelated genes of 8,584 with 20 degrees of freedom was not significant (P = 0.9872). We therefore have no evidence of stratification in our data [Pritchard and Rosenberg, 1999]. Our use of cases and controls for each recruitment center also minimized the potential for stratification.

Multiple Testing

The consistency of the SERPINA1 SNP combinations that contribute to risk, combined with the strength of the association and the absence of any association with SERPINA3 haplotypes, make it very unlikely that these results were a spurious outcome of multiple testing.

DISCUSSION

We identified six haplotypes in SERPINA1 that were risk factors for COPD (Table 5). They included one 5-SNP, two 6-SNP, two 7-SNP, and one 8-SNP haplotypes, all of which have a common 5-SNP backbone. This is suggestive of a basic 5-SNP risk-associated
The haplotypes of SERPINA1 identified in this study occur at a frequency similar to that of the Z variant, but have higher odds ratios for disease than possession of the Z allele. These findings highlight the role of specific haplotypes of SERPINA1 that confer risk to COPD, and further emphasize the key role of SERPINA1 in this disease.

ACKNOWLEDGMENTS

We thank Ann Hann (Bristol), Breda Callaghan and Gemma Hogan (Dublin), Joyce Barr (Edinburgh), and Clara Kolster-Bijdevaate (Leiden) for assisting with recruitment; G. Catapano, E. Fornai, and C. Carli for clinical and technical assistance in Pisa; Erika Daly (Dublin) for assistance in data analysis; and Drs. B. Veldhuisen and J.J. Houwing (Leiden) for comments on data analysis.

REFERENCES


