Genome-Wide Significant Association Between a Sequence Variant at 15q15.2 and Lung Cancer Risk

Thorunn Rafnar1, Patrick Sulem2, Soren Besenbacher1, Daniel F. Gudbjartsson1, Carlo Zanon1, Julius Gudmundsson1, Simon N. Stacey1, Jelena P. Kostic1, Thorgeir E. Thorgeirsson1, Gudmar Thorleifsson1, Hjordis Bjarnason1, Halla Skuladottir2, Tomas Gudbjartsson3,6, Helgi J. Isaaksson4, Dolores Isla7, Laura Murillo7, Maria D. García-Prats8, Angeles Panadero9, Katja K.H. Aben10,11, Sita H. Vermeulen11, Henricus F.M. van der Heijden12, William J. Feser14, York E. Miller18, Paul A. Bunn15, Augustine Kong1, Holly J. Wolf17, Wilbur A. Franklin16, Jose I. Mayordomo7, Lambertus A. Kiemeney10,11,13, Steinn Jonsson5,6, Unnur Thorsteinsdottir1,6, and Kari Stefansson1,6

Abstract

Genome-wide association studies (GWAS) have identified 3 genomic regions, at 15q24–25.1, 5p15.33, and 6p21.33, which associate with the risk of lung cancer. Large meta-analyses of GWA data have failed to find additional associations of genome-wide significance. In this study, we sought to confirm 7 variants with suggestive association to lung cancer (P < 10^-8) in a recently published meta-analysis. In a GWA dataset of 1,447 lung cancer cases and 36,256 controls in Iceland, 3 correlated variants on 15q15.2 (rs50417, rs11853991, and rs748404) showed a significant association with lung cancer, whereas rs4254535 on 2p14, rs530057 on 3p24.1, rs6433847 on 3q13.31, and rs1926203 on 10q23.31 did not. The most significant variant, rs748404, was genotyped in an additional 1,299 lung cancer cases and 4,102 controls from the Netherlands, Spain, and the United States and the results combined with published GWAS data. In this analysis, the T allele of rs748404 reached genome-wide significance (OR = 1.15, P = 1.1 × 10^-9). Another variant at the same locus, rs12050604, showed association with lung cancer (OR = 1.09, 3.6 × 10^-8) and remained significant after adjustment for rs748404 and vice versa. rs748404 is located 140 kb centromeric of the TP53BP1 gene that has been implicated in lung cancer risk. Two fully correlated, nonsynonymous coding variants in TP53BP1, rs260214 (Q1136K) and rs560191 (E353D) showed association with lung cancer in our sample set; however, this association did not remain significant after adjustment for rs748404. Our data show that 1 or more lung cancer risk variants of genome-wide significance and distinct from the coding variants in TP53BP1 are located at 15q15.2. Cancer Res; 71(4); 1356–61. © 2011 AACR.

Introduction

Lung cancer is the most common form of cancer and causes more deaths worldwide than any other cancer (1). The single most important factor influencing the risk of lung cancer is smoking. However, studies on lung cancer risk in family members of lung cancer probands have shown that genetic factors also contribute to the risk of developing the disease (2, 3). Furthermore, although smoking is the primary cause of lung cancer, only about 15% of lifelong smokers develop the disease, pointing to a role for genetic susceptibility (4, 5).

In the last 2 years, genome-wide association studies (GWAS) have yielded variants at 3 loci that associate with lung cancer in populations of European descent. First, sequence variants within a cluster of nicotinic acetylcholine receptor genes on chromosome 15q25 were simultaneously shown to associate with the number of cigarettes smoked per day (6), nicotine dependence (6, 7), and smoking-related diseases such as lung cancer (6, 8, 9) and peripheral arterial disease (6). The other 2 loci are at 5p15.33 within the TERT and CLPTM1L genes (10–12) and 6p21.33 within the MHC locus (8, 11). Neither locus is associated with smoking behavior.

Recently, Broderick et al. reported a meta-analysis of lung cancer based on 7,560 cases and 8,205 controls, genotyped for 21,620 single-nucleotide polymorphisms (SNP) (13). Apart from variants at the 3 previously reported loci, no variants...
showed a significant association with lung cancer after correcting for the number of tests ($P < 1 \times 10^{-5}$). Seven SNPs showed suggestive associations with $P$ values less than $10^{-4}$. The objective of this study was to test these variants in several independent lung cancer case–control sample sets to determine if they could be confirmed as lung-cancer-associated variants of genome-wide significance.

Materials and Methods

Study populations (summarized in Table 1)

**Iceland.** The Icelandic lung cancer study population was previously described (6). In brief, information on all lung cancer cases diagnosed from 1955 to 2008 was obtained from the Icelandic Cancer Registry. Samples from 848 cases were genotyped on the Human Hap300 or HumanCNV370-duo Bead Arrays. In addition to using data from directly genotyped cases, we used a method where genotypes of relatives are used to provide information on lung cancer cases not genotyped (familial imputation; ref 14). Using genotypes from relatives of 2,172 ungenotyped lung cancer cases, it was possible to infer genotypes that are equivalent to 639 lung cancer cases. The 36,256 controls used in this study consisted of individuals from other ongoing GWAS at deCODE that have been genotyped with the Human Hap300 or HumanCNV370-duo Bead Arrays. No individual disease group accounts for more than 6% of the total control group. The group of 15,310 smokers were a part of a study on smoking behavior and nicotine dependence that is described in detail in a recent publication (15). Study protocols were approved by the National Bioethics Committee of Iceland and all subjects gave written informed consent.

**Nijmegen, the Netherlands.** The Dutch study population was previously described (12). The 552 patients with lung cancer were identified through the population-based cancer registry of the Comprehensive Cancer Center IKO, Nijmegen, the Netherlands, and recruited through several independent studies (12). All participants gave informed consent for DNA-related research and linkage with disease registries. The 1,832 cancer-free controls (46% males) were selected from participants of the “Nijmegen Biomedical Study” (NBS; ref 16). All controls are of self-reported European descent. The study protocols of the NBS were approved by the Institutional Review Board of the RUNMC and all study subjects signed a written informed consent form.

**Zaragoza, Spain.** The Spanish study population was previously described (12). The 548 lung cancer cases were recruited from the Oncology Department of Zaragoza Hospital in Zaragoza, from June 2006 to June 2009. The 1,432 Spanish controls attended the University Hospital in Zaragoza for diseases other than cancer between November 2001 and May 2007. All patients and controls are of self-reported European descent. Study protocols were approved by the Institutional Review Board of Zaragoza University Hospital. All subjects gave written informed consent.

**Denver, Colorado.** DNA samples from blood and clinical data were provided by the University of Colorado Cancer Center under COMIRB protocol 08-0380. Blood samples were collected from 1,217 patients diagnosed with different diseases enrolled in any of 20 clinical research trials carried out at Colorado SPORE protocols between 1993 and 2008. Of these 1,217 patients, 246 were lung cancer cases and 971 had never had lung cancer at the time of sample shipment. Lung cancer cases were identified either from data matches with the Colorado Central Cancer Registry or by having malignant lung tissue collected via enrollment in a surgical protocol. Work in this study was limited to cases and controls of self-reported European ancestry (196 cases and 838 controls).

Data from published GWAS studies. In addition to the sample sets described above, we used association results from previously published studies. The study by Broderick et al., from which we selected the 7 SNPs to follow-up, uses sample sets from UK, IARC, and Texas and does not overlap with the populations directly genotyped for this study (Iceland, the Netherlands, Spain, and USA-Denver; ref 13). The study by Landi et al., contains data from 13,300 cases and 19,666 controls and this dataset partly overlaps with the Icelandic sample set (17). To use only nonoverlapping data when performing meta-analysis for rs12050604, we combined the data presented by Landi et al. only with data from the Netherlands, Spain, and USA-Denver in the combined analysis.
Genotyping

The Icelandic and cases and controls and the Dutch controls were assayed using the Human Hap300 and HumanCNV370-duo Bead Arrays (Illumina; ref 6). All other samples, that is, the Dutch cases and the cases and controls from Spain and the United States were genotyped using the Centaurus (Nanogen) platform (18). The quality of each Centaurus SNP assay was evaluated by genotyping each assay in the CEU HapMap samples and comparing the results with the HapMap publicly released data. Assays with more than 1.5% mismatch rate were not used, and a linkage disequilibrium (LD) test was used for markers known to be in LD. Approximately 10% of the Icelandic cases that were genotyped on the Illumina platform were also genotyped using the Centaurus assays and the observed mismatch rate was less than 0.5%.

Statistical analysis

A likelihood procedure implemented in the NEMO software was used for the association analyses (19). We tested the association of an allele to cancer using a standard likelihood ratio statistic that, if the subjects were unrelated, would have asymptotically a \( \chi^2 \) distribution with 1 degree of freedom under the null hypothesis. Allelic frequencies rather than carrier frequencies are presented for the markers in the main text. Allele-specific ORs and associated P values were calculated assuming a multiplicative model for the 2 chromosomes of an individual. To adjust for possible population stratification and the relatedness amongst the Icelandic study subjects, we divided the \( \chi^2 \) test statistics from the whole-genome lung cancer scan using the method of genomic control (20), that is, the 304,073 \( \chi^2 \) test statistics were divided by their means, which were 1.06 for the Icelandic subjects that were chip genotyped and 1.10 for the combined dataset of genotyped and ungenotyped subjects whose genotypes were derived by familial imputation. The sample sets from the Netherlands, Spain, and the United States used in this study were recruited on a population basis and assumed to be unrelated. Results from multiple case–control groups were combined using a Mantel–Haenszel model in which the groups were allowed to have different population frequencies for alleles, haplotypes, and genotypes but were assumed to have common relative risks (21). This method was used both for combining data within study groups genotyped by us and to combine those results with published results (allelic ORs and \( P \) values) from Broderick et al. (13) and Landi et al. (17). All reported \( P \) values are two-sided.

Results

We assessed the 7 SNPs reported by Broderick et al. to have a suggestive association with lung cancer (\( P < 10^{-5} \)) in our independent GWAS dataset corresponding to 1,447 lung cancer cases and 36,256 controls from Iceland. Only the 3 SNPs on 15q15.2 (rs504417, rs11853991, and rs748404) showed an effect in the same direction as previously reported with an association with lung cancer that remained significant for all 3 SNPs when taking the 7 tests into account (\( P < 0.05/7 = 0.007; \) Table 2). The SNPs are highly correlated \( (\text{r}^2 > 0.8) \) and in agreement with the results of Broderick et al., the strongest association was to the T allele of rs748404 \( (O R = 1.20, P = 5.2 \times 10^{-8}) \). The other 4 SNPs, located on chromosomes 2p14, 3p24.1, 3q13.31, and 10q23.31, did not show significant association with lung cancer (all \( P \geq 0.1; \) Table 2).

To further explore the association between rs748404-T and lung cancer, we genotyped the SNP in sample sets from the Netherlands, Spain, and the United States (Table 1). In combined analysis of the Icelandic and follow-up sample sets, the association became more significant \( (P = 2.5 \times 10^{-8}) \) and the effect was identical to the effect reported by Broderick et al. \( (O R = 1.15; \) Table 3). When all the data in our study (2,739 cases and 40,485 controls) were combined with the summary results of Broderick et al., using the Mantel–Haenszel model, the association of rs748404-T with lung cancer reached the commonly used threshold for genome-wide significance \( (P = 5 \times 10^{-8}) \) with a combined OR of 1.15 and a combined \( P \) value of \( 1.1 \times 10^{-3} \) (Table 3). We imputed all HapMap CEU (phase II) SNPs in a 450 kb region centered on rs748404 into the chip-genotyped Icelandic samples (808 cases and 36,256 controls) and repeated the association analysis. Of the 201 SNPs

<table>
<thead>
<tr>
<th>SNP (allele)</th>
<th>CHR</th>
<th>Cases (N)</th>
<th>Cont. (N)</th>
<th>Cont. (F)</th>
<th>OR 95% CI</th>
<th>P</th>
<th>OR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4254535 (T)</td>
<td>2p14</td>
<td>1,452</td>
<td>36203</td>
<td>0.68</td>
<td>0.97</td>
<td>0.89–1.05</td>
<td>0.45</td>
</tr>
<tr>
<td>rs1530057 (A)</td>
<td>3p24.1</td>
<td>1,429</td>
<td>36224</td>
<td>0.048</td>
<td>1.01</td>
<td>0.83–1.23</td>
<td>0.93</td>
</tr>
<tr>
<td>rs6483437 (G)</td>
<td>3q13.31</td>
<td>1,454</td>
<td>36259</td>
<td>0.223</td>
<td>0.92</td>
<td>0.83–1.02</td>
<td>0.10</td>
</tr>
<tr>
<td>rs1926203 (T)</td>
<td>10q23.31</td>
<td>1,466</td>
<td>36273</td>
<td>0.674</td>
<td>0.99</td>
<td>0.90–1.08</td>
<td>0.76</td>
</tr>
<tr>
<td>rs504417 (G)</td>
<td>15q15.2</td>
<td>1,447</td>
<td>36228</td>
<td>0.301</td>
<td>0.88</td>
<td>0.80–0.96</td>
<td>0.0043</td>
</tr>
<tr>
<td>rs11853991 (G)</td>
<td>15q15.2</td>
<td>1,451</td>
<td>36246</td>
<td>0.301</td>
<td>0.88</td>
<td>0.80–0.96</td>
<td>0.0036</td>
</tr>
<tr>
<td>rs748404 (T)</td>
<td>15q15.2</td>
<td>1,447</td>
<td>36256</td>
<td>0.797</td>
<td>1.20</td>
<td>1.08–1.33</td>
<td>0.00052</td>
</tr>
</tbody>
</table>

Shown are the SNP name, allele tested, chromosome location, the number of individuals (N) and the allelic frequency (F) of the variant in cases and controls, allelic odds-ratio (OR), 95% CI, and \( P \) values based on the multiplicative model and the OR reported by Broderick et al. (ref. 13; OR*). All \( P \) values are two-sided.
imputed in the region, no SNP showed stronger association with lung cancer than rs748404. rs748404-T was not associated with gender ($P = 0.64$) or age at diagnosis ($P = 0.69$) among lung cancer cases. Furthermore, we found no association between rs748404 and smoking quantity as measured by cigarettes smoked per day in a large group of Icelandic smokers ($N = 13,510$, $P = 0.2$), indicating that the variant is not likely to be acting through smoking quantity. We analyzed the effect of rs748404 in different histological types of lung cancer versus controls and observed a significant association between the T allele and the risk of adenocarcinoma ($N = 6,750$, OR $= 1.16$, $P = 0.0095$) and squamous cell carcinoma ($N = 672$, OR $= 1.16$, $P = 0.030$; Supplementary Table 1). The number of cases with other histological types of lung cancer was too small ($N < 350$) to draw definitive conclusions.

Landi et al. reported the results of a meta-analysis of over 13,000 cases and 19,000 controls that included samples from the Icelandic sample set used here (719 cases and 6,030 controls), and the UK sample set used by Broderick et al. (1,978 cases and 1,438 controls; ref 17). In this meta-analysis, 5 SNPs at 15q15.2 had suggestive association with lung cancer ($P$ values ranging from $3.5 \times 10^{-4}$ to $4.2 \times 10^{-5}$). The association between these 5 SNPs and lung cancer in the Icelandic chip-genotyped individuals, a part of which was included in the dataset used by Landi et al., is shown in Supplementary Table 2. The 5 SNPs are correlated with rs748404 with $r^2$ ranging from 0.13 to 0.69 (Supplementary Table 3). Of the 5 SNPs, the 2 fully correlated variants rs2277532 and rs12050604 ($r^2 = 1$ in CEU) showed the strongest association with lung cancer in the report by Landi et al. (OR = 1.11; $P = 3.5 \times 10^{-6}$; ref 17). Those 2 SNPs are also the least correlated to rs748404 ($r^2 < 0.14$), potentially suggesting a second signal at this locus. We genotyped rs12050604 in the samples from the United States, the Netherlands, and Spain and combined these data with the published summary results from Landi et al., giving us a total of 14,480 cases and 22,824 controls (Supplementary Table 4). In this large dataset, the C allele of rs12050604 had an OR of 1.09 and a $P$ value of $3.6 \times 10^{-6}$, not reaching genome-wide significance.

The variant rs748404 is located between 2 transglutaminases genes, TGM5 and TGM7, and approximately 200 kb telomeric to rs12050604 that is in an intron of the E3 ubiquitin-protein ligase $UBR1$ gene (Fig. 1). The 2 variants are separated by a recombination hotspot and the $r^2$ between them is only 0.126. When we adjust the results for each SNP in a logistic regression, using the other SNP as a covariate and only including individuals who we genotyped for both markers (Iceland, the Netherlands, Spain, and the United States), results for rs748404 remain significant after adjusting for rs12050604 (OR = 1.09, $P = 0.037$) and rs12050604 remains significant after adjusting for rs748404 (OR = 1.12, $P = 0.0028$; Supplementary Table 5). Overall, these results indicate that neither rs748404 nor rs12050604 can, by themselves, fully account for the association observed between sequence variants in this region and lung cancer. This observation suggests that either a unique variant, capturing the effects of both rs748404 and rs12050604, remains to be discovered or that the region contains more than 1 variant that predisposes to lung cancer.
The variant that shows a genome-wide significant association with lung cancer, rs748404, is located at the beginning of a 750 kb LD block containing several genes (Fig. 1). Of those, the most notable with respect to lung cancer is TP53BP1, a DNA damage checkpoint protein that is located 140 kb telomeric to rs748404. TP53BP1 contains 3 common SNPs [minor allele frequency (MAF) > 5% in HapMap CEU] that encode missense mutations: rs2602141 (Q1136K), rs560191 (E353D), and rs689647 (S412G). rs2602141 and rs560191 are fully correlated ($r^2 = 1$) and the common alleles of these SNPs have previously been reported to be associated with risk of lung cancer (22, 23). In our Icelandic dataset of 1,417 cases and 36,256 controls, the T allele of rs2602141 was associated with lung cancer (OR = 1.11, P = 0.011). We genotyped rs2602141 in our sample sets from the Netherlands, Spain, and the United States and combined the data with the Icelandic data, as well as with data from previous candidate gene studies by Rudd et al. (22) and Truong et al. (23). This analysis shows that the missense variant rs2602141-T (and consequently, the fully correlated missense variant rs560191-G) is significantly associated with risk of lung cancer (OR = 1.11, P = 8.60 $\times 10^{-5}$; Supplementary Table 6). The variant rs748404 is moderately correlated with rs2602141 and rs560191 ($r^2 = 0.578$). By adjusting the results for the coding SNP rs2602141 with rs748404 and vice versa, using only individuals we genotyped for both markers (a total of 2,239 cases and 40,418 controls), we found that results for rs748404 remain significant after adjustment for rs2602141 (OR = 1.12, P = 0.027), whereas results for rs2602141 do not remain significant after adjustment for rs748404 (OR = 1.02, P = 0.74; Supplementary Table 7). Thus, there is no evidence for an association signal from the coding variants, rs2602141 and rs560191, that is not captured by rs748404. The 3rd missense mutation in TP53BP1, rs689647-T, did not show significant association with lung cancer in a group of 819 lung cancer cases and 10,724 controls from the Icelandic population (OR = 0.94, P = 0.53).

Discussion

In this study, we have confirmed that a region on 15q15.2 is the 4th lung cancer risk region discovered through GWA studies in Europeans. By combining several independent datasets, the association of rs748404 reaches genome-wide significance; however, our results suggest that rs748404 does not fully explain the association observed at the locus and that either there may be a yet-unidentified variant or that there may be more than 1 variant at this locus that associates with lung cancer risk. In this regard, the locus has similarities to the TERT locus on 5p15.33 that has been shown to contain several independent variants that associate with lung cancer (10–12). Furthermore, we observed that although the missense mutation in TP53BP1, rs2602141, is significantly associated with lung cancer, this association no longer remains when adjusted for rs748404.

We noted that rs748404 was not among the top 200 SNPs (all $P < 8 \times 10^{-5}$) reported by Landi et al. in the meta-analysis of over 13,000 lung cancer cases and 19,000 controls, although a SNP with an OR = 1.15 and MAF ~0.77 should have a probability of ~80% to be included on this list. However, the list of top 200 SNPs in the meta-analysis includes 2 SNPs that have an $r^2$ of 0.69 with rs748404 and P values of 2 $\times 10^{-9}$, supporting the conclusion that rs748404 or correlated SNPs are associated with lung cancer.

We are aware that we do not currently know all the variants that exist at the 15q15.2 locus. Comprehensive genomic sequencing will eventually allow a more complete dissection of the association between variants in the region and lung cancer risk. Currently, whole genome sequencing efforts, such as the 1,000 Genomes (1,000 G) project (24), are leading to the discovery of a large number of variants that will be imputed into large datasets for association analysis. However, it may not be possible to impute all these variants with certainty, requiring direct genotyping on large datasets to test if they associate with defined phenotypes.
Mapping of lung cancer risk regions is an important step toward understanding of the pathogenesis of lung cancer. Lung cancer risk variants could fall into at least 3 categories, that is, variants that affect risk of lung cancer regardless of smoking status, variants that increase vulnerability to the harmful effects of smoking in smokers and variants that affect smoking behavior. For this last category, in addition to the variants at 15q25, large meta-analyses on smoking behavior have yielded sequence variants at 8p11 and 19q13 that show genome-wide significant association with smoking quantity and nominally significant association with lung cancer (6, 15). Large scale collaborative efforts will be required to uncover new variants that associate with lung cancer and refine the association signals already discovered.

Disclosure of Potential Conflicts of Interest

T. Rafnar, P. Sulem, S. Beesenbacher, D.F. Gudbjartsson, C. Zanon, J. Gudmundsson, S.N. Stacey, J.P. Kostic, T.E. Thorgeirsson, G. Thorleifsson, H. Bjarnason, A. Kong, U. Thorsteinsdottir, and K. Stefansson are employees of deCODE genomics, a biotechnology company that intends to incorporate the variant described in this paper into its genetic testing services. The other authors disclosed no potential conflicts of interest.

Acknowledgments

We thank the individuals that participated in the study and whose contribution made this work possible. We also thank the nurses at deCODE’s participant recruitment center and the personnel at deCODE’s core facilities. We acknowledge the Icelandic Cancer Registry for assistance in the ascertainment of the Icelandic cancer patients.

Grant Support

This work was supported by European Union (EU) FP7-MC-IAPP Grant agreement no. 218071 (CancerGene), the EU grant GENADICT1: LSHM-CT-2004-005166 and SPORE in Lung Cancer grant (P50 CA058187) and EDRN grant CA85070 from the National Cancer Institute. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 9, 2010; revised November 16, 2010; accepted November 27, 2010; published OnlineFirst February 8, 2011.

References

17. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 9, 2010; revised November 16, 2010; accepted November 27, 2010; published OnlineFirst February 8, 2011.

References

17. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 9, 2010; revised November 16, 2010; accepted November 27, 2010; published OnlineFirst February 8, 2011.
Genome-Wide Significant Association Between a Sequence Variant at 15q15.2 and Lung Cancer Risk

Thorunn Rafnar, Patrick Sulem, Soren Besenbacher, et al.

Cancer Res  Published OnlineFirst February 8, 2011.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-10-2852

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2011/02/07/0008-5472.CAN-10-2852.DC1

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.