A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium


1International Agency for Research on Cancer (IARC), Lyon, France, 2Institute of Carcinogenesis, Cancer Research Centre, Moscow, Russia, 3Department of Epidemiology, Institute of Occupational Medicine, Lodz, Poland, 4The M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland, 5National Institute of Environmental Health, Budapest, Hungary, 6Regional Authority of Public Health, Banska Bystrica, Slovakia, 7Institute of Public Health, Bucharest, Romania, 8Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic, 9Palacky University, Olomouc, Czech Republic, 10Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic, 11Department of Hygiene, Epidemiology, and Medical Statistics, University of Athens School of Medicine, Athens, Greece, 12Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America, 13INSERM U946, Paris, France, 14CNRS UMR2020, Gustave Roussy Institute, Villejuif, France, 15Geneva Cancer Registry, Institute for Social and Preventive Medicine, University of Geneva, Geneva, Switzerland, 16Bremen Institute for Prevention Research and Social Medicine (BIPS), University of Bremen, Bremen, Germany, 17Unit of Cancer Epidemiology, University of Turin, Turin, Italy, 18National Cancer Institute, IRCCS, Aviano, Italy, 19General Hospital of Pordenone, Pordenone, Italy, 20Cancer Registry of Norway, Oslo, Norway, 21School of Medicine and Dentistry, University of Aberdeen, Aberdeen, United Kingdom, 22Department of Environmental Medicine and Public Health, University of Padova, Padova, Italy, 23Respiratory Epidemiology and Public Health, National Heart and Lung Institute, Imperial College, London, United Kingdom, 24Institut Català d’Oncologia (ICO), Barcelona, Spain, 25CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain, 26University of Newcastle Dental School, Newcastle, United Kingdom, 27University of Glasgow Dental School, Glasgow, Scotland, 28University of Leeds Centre for Epidemiology and Biostatistics, Leeds, United Kingdom, 29NHS NSS ISD, Edinburgh, Scotland, 30Trinity College School of Dental Science, Dublin, Ireland, 31Croatian National Cancer Registry, Croatian National Institute of Public Health, Zagreb, Croatia, 32National School of Public Health/FIOCRUZ, Rio de Janeiro, Brazil, 33Universidade Federal de Pelotas, Pelotas, Brazil, 34Universidade de Sao Paulo, Sao Paulo, Brazil, 35Institute of Oncology and Radiobiology, Havana, Cuba, 36Institute of Hygiene, Università Cattolica del Sacro Cuore, Rome, Italy, 37IRCCS San Raffaele Pisana, Rome, Italy, 38Gilling School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States of America, 39School of Medicine, University of North Carolina, Chapel Hill, North Carolina, United States of America, 40Pomeranian Medical University, Department of Genetics and Pathomorphology, International Hereditary Cancer Center, Szczecin, Poland, 41Pomeranian Medical University, Department of Hygiene, Epidemiology, and Public Health, Szczecin, Poland, 42Fred Hutchinson Cancer Research Centre, Seattle, Washington, United States of America, 43Penn State College of Medicine, Hershey, Pennsylvania, United States of America, 44University of California Los Angeles School of Public Health, Los Angeles, California, United States of America, 45University of Texas M. D. Anderson Cancer Center, Houston, Texas, United States of America, 46Instituto de Investigación Epidemiológica, San José, Costa Rica, 47Brown University, Providence, Rhode Island, United States of America, 48Boston University School of Public Health, Boston, Massachusetts, United States of America, 49Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota, United States of America, 50University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, 51Department of Otorhinolaryngology and Head and Neck Surgery, Maastricht University Medical Centre, Maastricht, The Netherlands, 52Department of 
Introduction

560,000 cases of upper aerodigestive tract (UADT) cancers (encompassing of the oral cavity, pharynx, larynx and esophagus) are estimated to occur each year world-wide [1]. Exposure to alcohol and tobacco [1] are the major UADT cancer risk factors in Europe and the Americas, with infection with human papilloma-virus also playing an important role [2].

Elevated familial relative risks are consistently reported for UADT cancers [3-7]. While this implies that genetics contributes to UADT cancer susceptibility, the identity of the specific genes involved remains unclear. Studies of common genetic variation and UADT cancer susceptibility have mostly employed a candidate gene approach, with a particular focus on the genes that metabolize alcohol [8]. The metabolism of alcohol releases the carcinogen acetaldehyde as an intermediate [9]. As genetic variation in alcohol metabolism genes appears to influence their rate of function [10,11], variants that lead to a relative increase in exposure to acetaldehyde are expected to confer carriers to an increased risk of UADT cancers [12]. Consistent with this hypothesis, genetic variation in the alcohol dehydrogenase (ADH) 1B, and the aldehyde dehydrogenase 2 (ALDH2) genes in Asian populations have been associated with UADT cancer risk [8,12,13]. Three independent variants ADH1B, ADH7 and ADH1C have also been associated with UADT cancer risk in European populations [14]. Common genetic variation in additional genetic pathways have also been considered, although with some exceptions, such as DNA repair [15,16], the results have been inconsistent [3].
Author Summary

We have used a two-phased study approach to identify common genetic variation involved in susceptibility to upper aero-digestive tract cancer. Using Illumina HumanHap300 beadchips, 2,091 UADT cancer cases and 3,513 controls from two large European multi-centre UADT cancer studies, as well as 4,821 generic controls, were genotyped for a panel of 317,000 genetic variants that represent the majority of common genetic in the human genome. The 19 top-ranked variants were then studied in an additional series of 6,514 UADT cancer cases and 7,892 controls of European descent from an additional 13 UADT cancer studies. Five variants were significantly associated with UADT cancer risk after the completion of both stages, including three residing within the alcohol dehydrogenase genes (ADH1B, ADH1C, ADH7) that have been previously described. Two additional variants were found, one near the ALDH2 gene and a second variant located in HEL308, a DNA repair gene. These results implicate two variants 4q21 and 12q24 and further highlight three ADH variants UADT cancer susceptibility.

The candidate gene based studies have tested only a very small proportion of common human genetic variation in relation to UADT cancer risk. To further investigate common genetic variation and susceptibility to UADT cancers, we have performed a genome-wide association study within the International Head and Neck Cancer Epidemiology (INHANCE) consortium, comprising genome wide analysis of 2,091 UADT cancer cases and 8,334 controls and replication analysis of the nineteen top ranked variants in an independent series consisting of 6,514 UADT cancer cases and 7,892 controls from thirteen additional studies.

Results

Genome-wide results

After exclusion of suboptimal DNA based on QC criteria, data from 2,091 cases and 3,513 study specific controls and 4,821 generic controls were available for statistical analyses (Table S1) with 294,620 genetic variants. The overall results did not show a large deviation from what was expected by chance (λ = 1.07) (Figure 1). One genetic variant, rs971074, was strongly associated with UADT cancers (p = 1.0 × 10⁻⁶), rs971074 is positioned in the ADH7 locus on chromosome 4q23 and is highly correlated (r² = 1.0 CEU hapmap) with the SNP in ADH7, rs1573496, that we have described previously to be associated with UADT cancer risk [14]. Similarly, rs1789924, which is highly correlated (r² = 0.97 CEU hapmap) with ADH1C rs698, was also highly ranked (p = 2 × 10⁻⁶).

Variant selection for replication

For further analysis we selected the twenty top ranked genetic variants (including rs971074) for replication (Figure S1). These included those genetic variants in the discovery phase that achieved a p-value of ≤ 1 × 10⁻⁶ (12 variants) as well as nonsynonymous variants that achieved a p-value of ≤ 1 × 10⁻³ (3 additional variants). We also included variants that achieved a p-value of ≤ 5 × 10⁻⁷ when restricting the analysis to a specific UADT cancer site (1 variant), or heavy drinkers (1 variant) (Table 1). Only one variant from each high r² group (r² > 0.8) was included. We additionally included the non-synonymous ADH1B variant, rs1229984, that has been previously associated with UADT cancers [14] but not genotyped or tagged by a proxy variant on the HumanHap300 BeadChip. The association between the top ranked genetic variants selected for replication and UADT cancer was not sensitive to adjustment for population structure using principal component analysis, or exclusion generic controls (Table S2). rs1573496 was genotyped for replication as a proxy for rs971074 (r² = 1.00) and rs698 for rs1789924 (r² > 0.97) due to availability of Taqman assays. A TaqMan assay for rs12827056 could not be designed and no highly correlated (r² > 0.95) proxy genetic variant was available, hence further investigation was not possible.

Replication and combined results

Five genetic variants at three loci, 4q21, 4q23 and 12q24, were significantly associated with UADT cancer risk in the replication series (assuming Bonferroni correction for 19 comparisons or p ≤ 0.003, or p = 0.05 for previously described variants) or in the combined analysis (p-value of ≤ 5 × 10⁻⁷) (Table 1) (Figure S2). Using imputed genotypes across the 4q21, 4q23 and 12q24 regions based on Caucasian individuals from the HapMap consortium, we did not identify any variants more strongly associated with UADT cancer risk than the SNPs genotyped on the beadchips directly (Figure 2).

Two novel variant loci were identified. rs4767364 located at 12q24 (p_replication = 4 × 10⁻⁴; p_combined = 2 × 10⁻⁶) was one of

![Figure 1. Manhattan plot of the ARCAGE and CE UADT cancer GWAS discovery phase.](https://www.plosgenetics.org/doi/fig/10.1371/journal.pgen.1001333.g001)
### Table 1. Results from the UADT cancer genome-wide and replication analysis.

<table>
<thead>
<tr>
<th>Alleles</th>
<th>Discovery phase</th>
<th>Replication phase</th>
<th>Combined*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker</td>
<td>ref</td>
<td>rsn</td>
<td>OR</td>
</tr>
<tr>
<td>rs1229984</td>
<td>4q32</td>
<td>C</td>
<td>0.52</td>
</tr>
<tr>
<td>rs1970140</td>
<td>4q32</td>
<td>G</td>
<td>0.20</td>
</tr>
<tr>
<td>rs749961</td>
<td>4q21</td>
<td>A</td>
<td>1.31</td>
</tr>
<tr>
<td>rs1789346</td>
<td>4q28</td>
<td>AG</td>
<td>0.91</td>
</tr>
<tr>
<td>rs484870</td>
<td>19p13</td>
<td>A</td>
<td>0.72</td>
</tr>
<tr>
<td>rs970174</td>
<td>6p21</td>
<td>G</td>
<td>1.03</td>
</tr>
<tr>
<td>rs2299851</td>
<td>6p21</td>
<td>A</td>
<td>0.27</td>
</tr>
<tr>
<td>rs16837730</td>
<td>1p35</td>
<td>C</td>
<td>0.68</td>
</tr>
<tr>
<td>rs3810481</td>
<td>2p13</td>
<td>A</td>
<td>0.88</td>
</tr>
<tr>
<td>rs4831393</td>
<td>3p24</td>
<td>A</td>
<td>0.91</td>
</tr>
<tr>
<td>rs19800105</td>
<td>1q22</td>
<td>G</td>
<td>0.94</td>
</tr>
<tr>
<td>rs4571952</td>
<td>2q12</td>
<td>A</td>
<td>1.03</td>
</tr>
<tr>
<td>rs1789245</td>
<td>1q23</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>rs2102199</td>
<td>3p22</td>
<td>C</td>
<td>1.05</td>
</tr>
<tr>
<td>rs19800105</td>
<td>1q22</td>
<td>G</td>
<td>0.94</td>
</tr>
<tr>
<td>rs4571952</td>
<td>2q12</td>
<td>A</td>
<td>1.03</td>
</tr>
<tr>
<td>rs1789245</td>
<td>1q23</td>
<td>C</td>
<td>1.00</td>
</tr>
<tr>
<td>rs2102199</td>
<td>3p22</td>
<td>C</td>
<td>1.05</td>
</tr>
<tr>
<td>rs19800105</td>
<td>1q22</td>
<td>G</td>
<td>0.94</td>
</tr>
<tr>
<td>rs4571952</td>
<td>2q12</td>
<td>A</td>
<td>1.03</td>
</tr>
</tbody>
</table>

**A** Genome-Wide Association Study of UADT Cancers

*Including "generic" controls (methods) with the exception of rs1229984. Adjusted by sex, study.

*Replication phase excluded the SA Latin American study (Table 4) that had been published previously.

*Combined* analysis considered heavy drinkers only.

*Results from the UADT cancer genome-wide and replication analysis. Alleles Discovery phase a Replication phase b f Combined a

*Reason for exclusion of variant from discovery phase: ADH1B candidate gene.*

*Allele frequencies and replication phase p-values adjusted by sex, study.*

*rs1229984 and rs749961 were genotyped for analysis considering heavy drinkers only.*

*Adjusted by age, sex, study.*

*Analysis considered heavy drinkers only.*

*Abbreviations: OR, odds ratio; 95% CI, 95% confidence interval.*

*P-values from two-sided p-value.*

*P-values from two-sided p-value.*
Figure 2. Imputation and LD patterns. Imputation and LD patterns across the (a) 4q23 (ADH loci), (b) 12q24 (ALDH2), and (c) 4q21 (HEL308). Upper panel: Single marker association results for imputed (green) and directly genotyped variants (blue). Imputation performed on 2,091 cases and 3,513 study specific controls (excluded generic controls). After adjustment for the five variants that presented with replication, no variant had a
multiple highly correlated SNPs ($r^2 \geq 0.8$) that presented evidence for association in the GWAS stage. It is located in a LD region including multiple genes including the *aldehyde dehydrogenase 2 (ALDH2)* (Figure 2), another key gene in alcohol metabolism (Figure 2). In stratified analysis in the combined 8,744 UADT cancer cases and 11,982 controls (Table S3), the association was more pronounced in esophageal cancers compared to other UADT cancer subsites ($p$ heterogeneity $= 0.01$) and exhibited borderline heterogeneity when stratifying by alcohol use (Figure 3). Some heterogeneity was noted by when stratifying by country ($p = 0.004$), although there was no discernable geographic distribution that could explain this heterogeneity (data not shown).

We noted little evidence for association between alcohol consumption and rs4767364 (Table 2), nor was there evidence for association in the GWAS stage. It is located in a LD region that contains multiple genes (Figure 2), notably a second region of LD that contains multiple genes. Candidate genes include the *aldehyde dehydrogenase 2 (ALDH2)* (Figure 2), another key gene in alcohol metabolism. The minor allele carriers of *ADH2* variants were significantly associated in the independent replication series or after correction for multiple testing at a genome wide level in combined analysis ($\text{p} \leq 5 \times 10^{-7}$). The risk effects noted with all five variants were less prominent in the replication series when compared with the initial finding in the discovery series, consistent with the notion of “winner’s curse” [17]. In combination we estimate these 5 variants are likely to explain only a small proportion (approximately 4%) of the UADT cancer familial risk.

### Discussion

Five genetic variants at three loci, 4q23, 12q24 and 4q21, were significantly associated with UADT cancers in the independent replication series or after correction for multiple testing at a genome wide level in combined analysis ($\text{p} \leq 5 \times 10^{-7}$). The risk effects noted with all five variants were less prominent in the replication series when compared with the initial finding in the discovery series, consistent with the notion of “winner’s curse” [17]. In combination we estimate these 5 variants are likely to explain only a small proportion (approximately 4%) of the UADT cancer familial risk.

#### 12q24

The 12q24 variant, rs4767364, is positioned in an extended region of LD that contains multiple genes. Candidate genes include the *aldehyde dehydrogenase 2 (ALDH2)* (Figure 2), another key gene in alcohol metabolism. The minor allele carriers of *ADH2* variants were significantly associated in the independent replication series or after correction for multiple testing at a genome wide level in combined analysis ($\text{p} \leq 5 \times 10^{-7}$). The risk effects noted with all five variants were less prominent in the replication series when compared with the initial finding in the discovery series, consistent with the notion of “winner’s curse” [17]. In combination we estimate these 5 variants are likely to explain only a small proportion (approximately 4%) of the UADT cancer familial risk.

#### 4q21

The 4q21 variant significantly associated with UADT cancers was rs1494961 located (Table 1) 20 Mb proximal to the *ADH1* gene cluster. There is no LD between rs1494961 and the *ADHI* variant rs1573496 was observed never drinkers ($p = 0.03$). Among ever drinkers in this pooled analysis, the minor allele carriers of rs1229984 reported consuming less alcohol than non-carriers ($p = 3 \times 10^{-8}$). The rs1573496 minor allele carriers similarly noted to consume somewhat less alcohol ($p = 0.002$), while rs698 minor allele carriers consumed slightly more ($p = 0.05$) (Table 2). Adjustment for alcohol consumption made little difference to the risk estimates for UADT cancer with all three variants (Table S4).

### Association in African Americans

We additionally genotyped the five variants significantly associated with UADT cancer in 537 African American UADT cancer cases and 539 controls noting a significant association for the 12q24 variant rs4767364 ($\text{p} = 0.004$) (Table 3). Nevertheless, the smaller sample size and potential differences in genetic architecture between European and African American populations (both in terms of allele frequencies and LD structure) limits our ability to assess these five alleles in African-Americans.
variant and lung cancer ($p = 3 \times 10^{-4}$) (Figure 4) suggests that the causal variant maybe relevant for cancers influenced by tobacco consumption in general.

4q23

The top two ranked variants (rs1573496 and rs698 and correlated variants) from the GWAS stage we have previously associated with UADT cancer risk [14]. The association between these variants, and a third variant, rs1229984, not included in the Humanhap300 beadchip but genotyped here based on our previous findings [14], and UADT cancer was independently replicated in the additional UADT cases and controls presented here ($p = 1 \times 10^{-1}$, $1 \times 10^{-2}$, and $0.01$ for rs1573496, rs1229984 and rs698, respectively).

The combined sample series presented here, totaling 8,774 UADT cancer cases and 11,982 controls, allowed further

---

**Figure 3. Stratified analysis of 4 replicated SNPs located near alcohol metabolism genes.** Estimates for rs1229984 (ADH1B), rs1573496 (ADH7), rs1042758 (ADH1C) and rs4767364 (ALDH2) were derived from a log-additive genetic model. ORs were adjusted by age, sex, study and were derived from fixed effects models. “Generic” controls were not included in this analysis.

doi:10.1371/journal.pgen.1001333.g003
Table 2. Association between rs1229984, rs1573496, rs698, rs4767364, and drinking intensity in ever drinkers expressed as mean of ml of ethanol consumed per day.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Controls</th>
<th>UADT Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>mean</td>
<td>CI 95%</td>
</tr>
<tr>
<td>rs1229984 (ADH1B)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>14,518</td>
<td>35.06</td>
<td>33.32–36.80</td>
</tr>
<tr>
<td>CT, TT</td>
<td>1,323</td>
<td>22.85</td>
<td>18.73–26.98</td>
</tr>
<tr>
<td>p-trend</td>
<td>3.3 x 10^{-20}</td>
<td>5 x 10^{-12}</td>
<td></td>
</tr>
<tr>
<td>rs1573496 (ADH7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td>12,936</td>
<td>35.02</td>
<td>33.21–36.82</td>
</tr>
<tr>
<td>GC, CC</td>
<td>2,926</td>
<td>30.29</td>
<td>27.36–33.22</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.002</td>
<td>0.03</td>
<td>0.008</td>
</tr>
<tr>
<td>rs698 (ADH1C)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td>5,574</td>
<td>32.05</td>
<td>29.69–34.41</td>
</tr>
<tr>
<td>TC</td>
<td>6,748</td>
<td>32.51</td>
<td>30.29–34.74</td>
</tr>
<tr>
<td>CC</td>
<td>2,377</td>
<td>36.15</td>
<td>32.92–39.39</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.05</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>rs4767364 (ALDH2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td>7,232</td>
<td>33.78</td>
<td>31.66–35.90</td>
</tr>
<tr>
<td>GC, CC</td>
<td>6,297</td>
<td>33.91</td>
<td>31.68–34.74</td>
</tr>
<tr>
<td>AA</td>
<td>1,527</td>
<td>35.84</td>
<td>31.91–39.77</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.60</td>
<td>0.14</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Adjusted mean of ml per day were derived from ANOVA. P-trend were derived from a linear regression with log(ml of ethanol per day) as an outcome using a log-additive genetic model. All estimates were adjusted by sex, age, study, pack-years (and case/control status when appropriate).
doi:10.1371/journal.pgen.1001333.t002

Figure 4. Association between 4q21 variant (rs1494961) and UADT and lung cancers. ORs were adjusted by age, sex, study and were derived from fixed effects models. “Generic” controls were not included in this analysis.
doi:10.1371/journal.pgen.1001333.g004
controls included: 1,385 individuals from the 1958 birth cohort, generic controls to further increase statistical power. These generic controls from these two studies. We additionally included 4,983 UADT cancer cases (squamous cell carcinomas) and 4,090 quantity for genome-wide genotyping was available for 2,230 centers from 9 European countries. DNA of sufficient quality and case control study conducted by IARC from 2002 to 2005 in 12 centers from 5 countries; and the ARCAGE [14,34,38] (europe study [14,37,34] conducted from 2000 to 2002, in 6 Europe) multicentre UADT cancer case-control studies (Table 4), based multi-centre UADT cancer case-control studies (Table 4), those genotyped were restricted to French and 433 Norwegian controls genotyped by the Centre National Genotypage (CNG Evry France). We also included in our control series a separate group of 1,342 kidney cancer cases from the same centres as the central Europe study, inclusion or exclusion of these “controls” had no material effect on the results presented (Table S2). Both studies have been approved by local ethics committees as well as IARC IRB.

### Materials and Methods

#### Discovery phase study samples

Genome-wide genotyping was performed in two European based multi-centre UADT cancer case-control studies (Table 4), the International Agency for Research on Cancer (IARC) central europe study [14,37,34] conducted from 2000 to 2002, in 6 centers from 5 countries; and the ARCAGE [14,34,33] [Alcohol-Related Cancers and Genetic susceptibility in Europe] multicentre case control study conducted by IARC from 2002 to 2003 in 12 centers from 9 European countries. DNA of sufficient quality and quantity for genome-wide genotyping was available for 2,230 UADT cancer cases (squamous cell carcinomas) and 4,990 controls from these two studies. We additionally included 4,983 generic controls to further increase statistical power. These generic controls included: 1,385 individuals from the 1958 birth cohort, (Wellcome Trust case control consortium[39]) as well as 1,823 French and 433 Norwegian controls genotyped by the Centre National Genotypage (CNG Evry France). We also included in our control series a separate group of 1,342 kidney cancer cases from the same centres as the central Europe study, inclusion or exclusion of these “controls” had no material effect on the results presented (Table S2). Both studies have been approved by local ethics committees as well as IARC IRB.

### Genome-wide genotyping and quality control

The central Europe study and the ARCAGE study were genotyped using the Illumina Sentrix HumanHap300 BeadChip at the Centre d’Etude du Polymorphisme Humain (CEPH) and the CNG as described previously [34,40]. We conducted systematic quality control steps on the raw Illumina HumanHap300 genotyping data. Variants with a genotype call rate of less than 95% and also individuals where the overall genotype completion rate was less than 95% were excluded. We also conducted further exclusions where the genotype distribution clearly deviated from that expected by Hardy-Weinberg Equilibrium (HWE) among controls (p-value of less than 10⁻⁷) and where there were discrepancies between sex based genotype and reported sex, as well as individuals with unlikely heterozygosity rates across genetic variants on the X chromosome (Table S1). Those genotyped were restricted to individuals of self-reported European ethnicity. To further increase the ethnic homogeneity of the series, we used the program STRUCTURE [41] to identify individuals of mixed ethnicity. Using a subseries of 12,898 genetic variants from the HumanHap 300 BeadChip panel evenly distributed across the genome and in low linkage disequilibrium (LD) (r²<0.004) [42], we estimated the genetic profile of the study participants compared with individuals of known ethnic origins (the Caucasian, African and east-Asian individuals genotyped by the HapMap project). We excluded 34 individuals because of some evidence of ethnic admixture (Figure S3), indicating that the extent of admixture within the central Europe and ARCADE study centers is limited.

### Genome-wide statistical analysis

The association between each genetic variant and the disease risk was estimated by the odds ratio (OR) per allele and ninety-five percent confidence intervals (CI) using multivariate unconditional logistic regression assuming a log-additive genetic model with sex and country of recruitment included in the regression model as covariates. Results that obtained a level of significance of a two sided p<5×10⁻⁷ were considered significant at a genome wide
level [39]. All analyses were conducted using PLINK [43]. We also conducted analyses restricting to UADT cancer subtypes (oral/pharyngeal cancer, laryngeal cancer, esophageal cancer) and restricting to heavy (median) drinkers and heavy (median) smokers.

The potential for population stratification not accounted for by adjustment by country was also investigated by principal components analysis (PCA) undertaken with the EIGENSTRAT package [44] using 12,898 markers in low LD [42]. Adjustment for population stratification using the PCA was performed by including significant eigenvectors that were associated with case control status (p < 0.05) as covariates in the logistic regression.

Genotypes for genetic variants across 4q21, 4q23 and 12q21 not genotyped on the Illumina HumanHap300 BeadChip, but genotyped by the HAPMAP consortium, were imputed using the program MACH with phased genotypes from the CEU Hapmap genotyping as a scaffold. Unconditional logistic regression using posterior haplotype probabilities (haplotype dosages) from MACH were carried out using ProbABEL [45] including age, sex, and country of origin in the regression as covariates. Linkage Disequilibrium (LD) statistics (D' and r²) were calculated using Haploview [46].

**Replication study samples**

The replication series consisted of 6,514 UADT cancer cases (squamous cell carcinomas) and 7,892 controls from 13 UADT cancer case-control studies (Table 4). With the exception of the Szczecin case-control study [16], all studies were part of the

<table>
<thead>
<tr>
<th>Study Name</th>
<th>Study setting</th>
<th>Coordinating centre</th>
<th>Genotyping centre</th>
<th>Principal Investigators</th>
<th>UADT Subsites*</th>
<th>Control source</th>
<th>Cases</th>
<th>Controls</th>
<th>Total Cases</th>
<th>Total Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCAGE b</td>
<td>Europe - Multicentre</td>
<td>IARC</td>
<td>CNG</td>
<td>Boffetta/ Brennan</td>
<td>UADT</td>
<td>Hospital-based</td>
<td>1,422</td>
<td>1,503</td>
<td>1,368</td>
<td>1,313</td>
</tr>
<tr>
<td>Central Europe c</td>
<td>Europe - Multicentre</td>
<td>IARC</td>
<td>CNG</td>
<td>Boffetta/ Brennan</td>
<td>UADT</td>
<td>Hospital-based</td>
<td>808</td>
<td>2,587</td>
<td>723</td>
<td>2,200</td>
</tr>
</tbody>
</table>

Generic controls 4,821

**Table 4. The 15 UADT cancer studies participating in the genome-wide and replication analysis.**

- Including only individuals of self-reported European ancestry.
- Includes countries: Czech Republic, Greece, Italy, Norway, UK, Spain, Croatia, Germany, France.
- Includes countries: Romania, Poland, Russia, Slovakia, Czech Republic.
- For the three variants at 4q23, results have been published previously, in “replication” analysis for these variants, the SA study was excluded.
- UADT – Oral, pharynx, laryngeal, esophageal cancers, HN – Head and neck cancers Oral, pharynx, laryngeal cancers.

doi:10.1371/journal.pgen.1001333.t004

[39] 2011 | Volume 7 | Issue 3 | e1001333
The association between the nineteen variants and UADT cancer risk was estimated by per allele ORs and their 95% CI derived from multivariate unconditional logistic regression, with age, sex, and study (and country of origin where appropriate) included in the regression model as covariates. Measures of alcohol consumption have been previously harmonized across INHANCE studies [48]. The association between ADH/ALDH2 variants and alcohol consumption was carried out in ever drinkers using multivariate linear regression using a log transformed milliliter of ethanol consumed per day as an outcome, adjusting for age, sex, study, packyears (case-control status when appropriate). Milliliters of ethanol consumed per day was not available for 3 studies (Szczecin, Philadelphia/New York and The Netherlands study). Heterogeneity of ORs across the studies and across the stratification groups was assessed using the Cochran’s Q-test. All replication and combined analyses were conducted using SAS 9.1 software. P values were two sided.

Investigation of the effects of 4q21 variant rs1494961 and lung cancer risk

The series of lung cancer cases and controls used to investigate 4q21 variant, rs1494961, and lung cancer risk included studies from central Europe (IARC), Toronto (McGill), HUNT2/Tromso, the CARET cohort, EPIC-lung, the Szczecin case-control study, Liverpool Lung Project (LLP), Paris France and Estonia as described previously [34,40,49]. All studies have been approved by local ethics committees as well as IARC IRB.

Genotyping protocol for 4q21 variant, rs1494961

Genotyping for rs1494961 was performed using the Illumina beadchips (Central Europe [IARC], Toronto [McGill], HUNT2/Tromso, the CARET cohort, France and Estonia) or the Applied Biosystems Taqman assays (EPIC-lung, the Szczecin case-control study, Liverpool Lung Project [LLP] at IARC.

For the central European lung cancer study, the controls overlapped with the central European UADT cancer study for Bucharest (Romania), Lodz (Poland), Moscow (Russia), Banska Bystrika (Slovakia), and Olomouc and Prague (Czech Republic). We therefore performed analyses both including and excluding centres where controls overlapped.

Web resources

http://inhance.iarc.fr/ (December 2010)
http://www.hapmap.org (December 2010)
http://www.sph.umich.edu/csg/abecasis/mach/index.html (December 2010)

Supporting Information

Figure S1 Strategy for discovery and replication in the genome-wide association study.
Found at: doi:10.1371/journal.pgen.1001333.s001 (0.17 MB DOC)

Figure S2 Analysis of selected variants by study and by UADT cancer site in the replication series. For replication estimates of rs1229984, rs1573496, rs698, the SA study was excluded. Found at: doi:10.1371/journal.pgen.1001333.s002 (0.26 MB DOC)

Figure S3 STRUCTURE Admixture plots. Individuals plotted against individuals of known Caucasian (CEU), African (YRI) and East Asian (JPT-CHB) origin. Individuals with greater than 30% admixture (dashed line) were excluded. Found at: doi:10.1371/journal.pgen.1001333.s003 (0.30 MB DOC)

Table S1 Exclusion criteria of subjects for GWAS.
Found at: doi:10.1371/journal.pgen.1001333.s004 (0.18 MB DOC)

Table S2 Sensitivity analysis on the top variants identified by the genome-wide analysis.
Found at: doi:10.1371/journal.pgen.1001333.s005 (0.24 MB DOC)

Table S3 Selected demographic characteristics of cases and controls (GWAS and replication data combined).
Found at: doi:10.1371/journal.pgen.1001333.s006 (0.20 MB DOC)

Table S4 Comparison between analysis adjusted and unadjusted on tobacco and alcohol consumption.
Found at: doi:10.1371/journal.pgen.1001333.s007 (0.16 MB DOC)

Table S5 Minor allele frequency of each variant per study.
Found at: doi:10.1371/journal.pgen.1001333.s008 (0.22 MB DOC)

Acknowledgments

The authors thank all of the participants who took part in this research and the funders and support and technical staff who made this study possible. We thank Paul Pharoh and SEARCH for contribution of biological samples.
Author Contributions


References


