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. Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland, 5National Institute of Environmental Health, Budapest, Hungary, 6Regional Authority of Public Health, Banská 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, 13INSPER 1946, Paris, France, 14CNRS UMR8200, 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, IRCSS, 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...
Abstract

Genome-wide association studies (GWAS) have been successful in identifying common genetic variation involved in susceptibility to etiologically complex disease. We conducted a GWAS to identify common genetic variation involved in susceptibility to upper aerodigestive tract (UADT) cancers. Genome-wide genotyping was carried out using the Illumina HumanHap300 beadchips in 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. The 19 top-ranked variants were investigated further in an additional 6,514 UADT cancer cases and 7,892 controls of European descent from an additional 13 UADT cancer studies participating in the INHANCE consortium. Five common variants presented evidence for significant association in the combined analysis (p=5×10^{-7}). Two novel variants were identified, a 4q21 variant (rs1494961, p=1×10^{-8}) located near DNA repair related genes HEL308 and FAM175A (or Abaxas) and a 12q24 variant (rs4767364, p=2×10^{-7}) located in an extended linkage disequilibrium region that contains multiple genes including the aldehyde dehydrogenase 2 (ALDH2) gene. Three remaining variants are located in the ADH gene cluster and were identified previously in a candidate gene study involving some of these samples. The association between these three variants and UADT cancers was independently replicated in 5,092 UADT cancer cases and 6,794 controls non-overlapping samples presented here (rs1573496-ADH7, p=5×10^{-7}; rs1229984-ADH1B, p=7×10^{-9}; and rs698-ADH1C, p=0.02). These results implicate two variants at 4q21 and 12q24 and further highlight three ADH variants in UADT cancer susceptibility.

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 worldwide [1]. Exposure to alcohol and tobacco [1] are the major UADT cancer risk factors in Europe and the Americas, with infection with human papillomavirus 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 variants 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 Human-Hap300 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 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<1×10⁻⁸), 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 ≤1×10⁻³ (12 variants) as well as nonsynonymous variants that achieved a p-value of ≤1×10⁻⁵ (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 (preplication = 4×10⁻⁷; pcombined = 2×10⁻⁶) was one of
### Table 1. Results from the UADT cancer genome-wide and replication analysis.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chromosome region</th>
<th>Alleles</th>
<th>Discovery phase&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Replication phase&lt;sup&gt;bf&lt;/sup&gt;</th>
<th>Combined&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ref</td>
<td>rare</td>
<td>Reason for replication attempt</td>
<td>OR, 95% CI</td>
<td>OR, 95% CI</td>
</tr>
<tr>
<td>rs1229984</td>
<td>4q23</td>
<td>C</td>
<td>ADH1B, candidate gene</td>
<td>0.52, 0.43–0.64</td>
<td>0.64, 0.59–0.71</td>
</tr>
<tr>
<td>rs971074&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4q23</td>
<td>G</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>0.70, 0.62–0.79</td>
<td>0.78, 0.72–0.86</td>
</tr>
<tr>
<td>rs1494961</td>
<td>4q21</td>
<td>T</td>
<td>non-synonymous and p&lt;10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>1.15, 1.07–1.24</td>
<td>1.11, 1.06–1.17</td>
</tr>
<tr>
<td>rs4767364</td>
<td>12q24</td>
<td>G</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.21, 1.12–1.32</td>
<td>1.10, 1.04–1.15</td>
</tr>
<tr>
<td>rs1789924&lt;sup&gt;d&lt;/sup&gt;</td>
<td>4q23</td>
<td>T</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.20, 1.11–1.29</td>
<td>1.07&lt;sup&gt;f&lt;/sup&gt;, 1.01–1.14</td>
</tr>
<tr>
<td>rs1431918</td>
<td>8q12</td>
<td>G</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.19, 1.10–1.28</td>
<td>1.05, 1.00–1.11</td>
</tr>
<tr>
<td>rs7431530</td>
<td>3p24</td>
<td>C</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>0.81, 0.74–0.88</td>
<td>0.95, 0.90–1.00</td>
</tr>
<tr>
<td>rs3810481</td>
<td>20q13</td>
<td>G</td>
<td>non-synonymous and p&lt;10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>1.22, 1.11–1.34</td>
<td>1.07, 0.99–1.15</td>
</tr>
<tr>
<td>rs10801805</td>
<td>1p22</td>
<td>G</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.20, 1.11–1.29</td>
<td>1.04, 0.98–1.10</td>
</tr>
<tr>
<td>rs1041973</td>
<td>2q12</td>
<td>C</td>
<td>non-synonymous and p&lt;10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>0.83, 0.76–0.90</td>
<td>0.94, 0.89–1.00</td>
</tr>
<tr>
<td>rs4799863</td>
<td>18q12</td>
<td>A</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>0.84, 0.78–0.91</td>
<td>0.96, 0.92–1.01</td>
</tr>
<tr>
<td>rs2517452&lt;sup&gt;d&lt;/sup&gt;</td>
<td>6p21</td>
<td>C</td>
<td>p_oral&lt;5.10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>0.69, 0.59–0.80</td>
<td>1.00, 0.87–1.15</td>
</tr>
<tr>
<td>rs2012199</td>
<td>1q23</td>
<td>T</td>
<td>non-synonymous and p&lt;10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>1.24, 1.11–1.38</td>
<td>1.06, 0.98–1.14</td>
</tr>
<tr>
<td>rs2287802</td>
<td>19p13</td>
<td>A</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.19, 1.11–1.28</td>
<td>1.02, 0.97–1.07</td>
</tr>
<tr>
<td>rs1637730&lt;sup&gt;*&lt;/sup&gt;</td>
<td>1p35</td>
<td>C</td>
<td>p heavy drinkers &lt;5.10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>2.06, 1.57–2.71</td>
<td>1.02, 0.86–1.22</td>
</tr>
<tr>
<td>rs11067362</td>
<td>12q24</td>
<td>T</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.35, 1.19–1.54</td>
<td>1.01, 0.92–1.10</td>
</tr>
<tr>
<td>rs7924284</td>
<td>10q24</td>
<td>C</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>1.38, 1.20–1.59</td>
<td>1.00, 0.90–1.11</td>
</tr>
<tr>
<td>rs484870</td>
<td>19p13</td>
<td>A</td>
<td>non-synonymous and p&lt;10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>1.16, 1.08–1.26</td>
<td>0.99, 0.94–1.05</td>
</tr>
<tr>
<td>rs2299851</td>
<td>6p21</td>
<td>G</td>
<td>p_all&lt;10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>0.72, 0.62–0.83</td>
<td>1.11, 1.00–1.23</td>
</tr>
</tbody>
</table>

<sup>a</sup> Including “generic” controls (methods) with the exception of rs1229984. Adjusted by sex, study.
<sup>b</sup> Adjusted by age, sex, study.
<sup>c</sup> rs970174 and rs1789924 were genotyped in the replication phase by highly correlated variants (r<sup>2</sup> = 0.97) rs1573496 and rs698.
<sup>d</sup> Analysis considered oral cancers only.
<sup>e</sup> Analysis considered heavy drinkers only.
<sup>f</sup> For 4q23 variants rs1229984, rs1573496, rs698, the replication phase excluded the SA Latin American study (Table 4) that had been published previously.
P<sub>ts</sub>: two-sided p-value.

doi:10.1371/journal.pgen.1001333.t001
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 \textit{aldehyde dehydrogenase 2} (\textit{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 any gene-gene interactions between associated variants in \textit{ADH} gene cluster and rs4767364 (data not shown).

The second additional locus identified was at 4q21, with the nonsynonymous variant rs1494961 located in the \textit{HEL308} gene ($p_{\text{replication}} = 2 \times 10^{-5}$, $p_{\text{combined}} = 1 \times 10^{-5}$) (Table 1). In combined analysis, the association tended to be more pronounced in younger ages and smokers (Figure 4). Given the possible role of the \textit{HEL308} in DNA repair, we also investigated the possibility that rs1494961 may play a role in lung cancer susceptibility by genotyping rs1494961 in a series of 5,652 lung cancer cases and 9,388 controls. We noted a similar association between rs1494961 and lung cancer ($OR = 1.09$, $p = 3 \times 10^{-5}$) from nine lung cancer studies, even when we excluded 1,844 cases and 2,753 controls where controls overlapped with the central European UADT study ($OR = 1.09$, $p = 0.005$).

Replication of \textit{ADH} genes associations

The association between the \textit{ADH} variants, rs1573496, rs1229984 and rs698 at 4q23 and UADT cancer was described previously [14] in the CE, ARCAGE (excluding Bremen) and SA studies. When excluding these studies, the association with these variants was independently replicated in the additional 5,092 UADT cancer cases and 6,794 controls presented here ($p = 5 \times 10^{-6}$, $7 \times 10^{-3}$ and 0.02 for rs1573496, rs1229984 and rs698, respectively) (Table 1). We combined all studies totaling 8,744 UADT cancer cases and 11,982 controls to investigate effects of the \textit{ADH} variants among different strata (Figure 3). For both the \textit{ADH1B} and \textit{ADH7} variants heterogeneity was noted by UADT cancer subsite ($p_{\text{heterogeneity}} = 0.002$, and 0.06 respectively). The rs1229984 \textit{ADH1B} variant showed strong heterogeneity when stratifying by alcohol, with little evidence for association in never drinkers. By contrast, there was little evidence for heterogeneity noted with rs1573496 and rs698, but a statistically significant association with the \textit{ADH7} 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 ($p = 3 \times 10^{-5}$). rs1573496 minor allele carriers similarly were 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 ($p = 0.004$) (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.

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 ($p \leq 5 \times 10^{-7}$). 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 \textit{aldehyde dehydrogenase 2} (\textit{ALDH2}) (Figure 2), another key gene in alcohol metabolism. The minor allele carriers of \textit{ALDH2} variants rs737280 and rs4643328, in LD with rs4767364 ($r^2 = 0.86$ and 0.67, respectively), have been associated with differences in alcohol metabolism in Europeans, leading some authors to hypothesise [18] that these alleles have a similar, albeit more modest, effect in European populations to that of the \textit{ALDH2} rs671 variant linked to alcohol metabolism differences [10] in Asian populations. The increased UADT cancer risk we observed with the minor allele of rs4767364 (and rs737280 by imputation, Figure 2) is similar to the UADT cancer risk effect observed for heterozygote rs671 carriers [12,13] and is consistent with this hypothesis. Nevertheless, this region contains many additional plausible candidate genes. Other GWAS have implicated multiple variants in this region in many phenotypes (type 1 diabetes, arthritis, renal function, hemoglobin concentration/hematocrit, coronary artery disease and waist-to-torso ratio) [19-26] and therefore the nature of the actual causative allele and gene remains to be determined. The rs4767364 variant was also associated with UADT cancer risk in a smaller series of African Americans implying that this effect may be relevant to other populations.

4q21

The 4q21 variant significantly associated with UADT cancers was rs1494961 located (Table 1) 20 Mb proximal to the \textit{ADH} gene cluster. There is no LD between rs1494961 and either rs1229984, rs1573496 or rs698 ($r^2 < 0.003$). rs1494961 is a non-synonymous variant positioned in the \textit{HEL308} gene, a single stranded DNA-dependent ATPase and DNA helicase involved in DNA intra-strand cross-linking repair [27], although the residue involved, I306V, is not an evolutionary conserved site [28] suggesting that this alteration may not have a functional consequence. rs1494961 strand cross-linking repair [27], although the residue involved, I306V, is not an evolutionary conserved site [28] suggesting that this alteration may not have a functional consequence. rs1494961 is in a LD region spanning approximately 90 kb, and is highly correlated ($r^2 > 0.95$) with more than 20 common genetic variants. This region contains additional genes (Figure 2), notably a second DNA repair-related gene, \textit{FAM175A} (or \textit{Abraxas} and \textit{CCDC96}), that interacts directly with the BRCT repeat region of \textit{BRCAl} [29]. That a comparable association was noted between this
variant and lung cancer ($p = 3 \times 10^{-5}$) (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 ($p = 1 \times 10^{-5}$, $1 \times 10^{-6}$ 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 Controls</th>
<th>UADT Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n  mean CI 95%</td>
<td>n  mean CI 95%</td>
</tr>
<tr>
<td>rs1229984 (ADH1B)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>14,518 35.06 33.32–36.80</td>
<td>7,742 22.43 20.91–23.95</td>
</tr>
<tr>
<td>CT,TT</td>
<td>1,323 22.85 18.73–26.98</td>
<td>907 15.60 12.35–18.85</td>
</tr>
<tr>
<td>p-trend</td>
<td>$3 \times 10^{-20}$</td>
<td>$5 \times 10^{-10}$</td>
</tr>
<tr>
<td>rs1573496 (ADH7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td>12,936 35.02 33.21–36.82</td>
<td>6,823 22.36 20.79–23.93</td>
</tr>
<tr>
<td>GC, CC</td>
<td>2,926 30.29 27.36–33.22</td>
<td>1,822 19.68 17.25–22.11</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.002</td>
<td>0.03</td>
</tr>
<tr>
<td>rs698 (ADH1C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td>5,574 32.05 29.69–34.41</td>
<td>3,054 20.65 18.65–22.66</td>
</tr>
<tr>
<td>GC</td>
<td>2,926 30.29 27.36–33.22</td>
<td>1,822 19.68 17.25–22.11</td>
</tr>
<tr>
<td>CC</td>
<td>2,377 36.15 32.92–39.39</td>
<td>1,285 20.42 17.63–23.22</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.05</td>
<td>0.32</td>
</tr>
<tr>
<td>rs4767364 (ALDH2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td>7,232 33.78 31.66–35.90</td>
<td>4,114 21.40 19.57–23.22</td>
</tr>
<tr>
<td>GC</td>
<td>2,926 30.29 27.36–33.22</td>
<td>1,822 19.68 17.25–22.11</td>
</tr>
<tr>
<td>AA</td>
<td>1,527 35.84 31.91–39.77</td>
<td>718 25.09 21.43–28.74</td>
</tr>
<tr>
<td>p-trend</td>
<td>0.60</td>
<td>0.14</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. Estimates were derived from a log-additive 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.g004
A Genome-Wide Association Study of UADT Cancers

Table 3. Comparison of results from the genome-wide analysis with analysis in a UADT case-control series of African-American origin.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Combined GWA and replication analysis (European descent)</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAF</td>
<td>OR</td>
</tr>
<tr>
<td>rs1229984</td>
<td>0.06</td>
<td>0.64</td>
</tr>
<tr>
<td>rs1573496</td>
<td>0.11</td>
<td>0.75</td>
</tr>
<tr>
<td>rs698</td>
<td>0.38</td>
<td>1.12</td>
</tr>
<tr>
<td>rs4767364</td>
<td>0.29</td>
<td>1.13</td>
</tr>
<tr>
<td>rs1494961</td>
<td>0.49</td>
<td>1.12</td>
</tr>
</tbody>
</table>

Ca: number of cases; Co: number of controls.

Materials and Methods

Genome-wide genotyping and quality control
The central Europe study and the ARCADE 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 subspecies of 12,898 genetic variants from the HumanHap 300 BeadChip panel evenly distributed across the genome and in low linkage disequilibrium (LD) (r^2<0.004) [42], we estimated the ethnic 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 4. The 15 UADT cancer studies participating in the genome-wide and replication analysis.

<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>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS</td>
<td>Post GWAS Qc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>
<tr>
<td>Generic controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Replication

- **SA** Latin America - Multicentre
  - IARC: IARC
  - Genotyping centre: IARC
  - Principal Investigators: Boffetta/Brennan
  - UADT Subsites: UADT
  - Control source: Hospital-based
  - Cases: 1,422
  - Controls: 1,098
  - Includes countries: Czech Republic, Greece, Italy, Norway, UK, Spain, Croatia, Germany, France.

- **ARCGAGE - Bremen**
  - Bremen - Germany
  - Genotyping centre: Bremen Uni.
  - Principal Investigators: Ahrens
  - UADT Subsites: UADT
  - Control source: Hospital-based
  - Cases: 164
  - Controls: 190
  - Includes countries: Romania, Poland, Russia, Slovakia, Czech Republic.

- **Rome**
  - Rome - Italy
  - Genotyping centre: Uni. Rome
  - Principal Investigators: Boccia
  - UADT Subsites: Hospital-based
  - Cases: 251
  - Controls: 237
  - Includes countries: Romania, Poland, Russia, Slovakia, Czech Republic.

- **Poland**
  - Szczecin - Poland
  - Genotyping centre: Szczecin Uni.
  - Principal Investigators: Lubinski
  - UADT Subsites: Hospital-based
  - Cases: 409
  - Controls: 1,039
  - For the three variants at 4q23, results have been published previously, in “replication” analysis for these variants, the SA study was excluded.

- **Seattle (Oral Gen study)**
  - Washington - US
  - Genotyping centre: Fred Hutchinson Cancer Research Centre
  - Principal Investigators: Schwartz/Chen
  - UADT Subsites: Hospital-based
  - Cases: 193
  - Controls: 388
  - Includes countries: Romania, Poland, Russia, Slovakia, Czech Republic.

- **University of North Carolina (CHANCE study)**
  - North Carolina - US
  - Genotyping centre: University of North Carolina
  - Principal Investigators: Olishan
  - UADT Subsites: Population-based
  - Cases: 940
  - Controls: 1,087
  - Includes countries: Romania, Poland, Russia, Slovakia, Czech Republic.

- **Penn State**
  - Philadelphia, New York City - US
  - Genotyping centre: Penn State University
  - Principal Investigators: Muscat/Lazarus
  - UADT Subsites: Hospital-based
  - Cases: 310
  - Controls: 534
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **UCLA**
  - Los Angeles - US
  - Genotyping centre: University of California, LA
  - Principal Investigators: Zhang
  - UADT Subsites: Hospital-based
  - Cases: 206
  - Controls: 577
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **MD Anderson**
  - Houston - US
  - Genotyping centre: MD Anderson Cancer Centre
  - Principal Investigators: Wei/Sturgis
  - UADT Subsites: Hospital-based
  - Cases: 431
  - Controls: 431
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **IARC - oral cancer (ORC)**
  - Europe - Multicentre
  - Genotyping centre: IARC
  - Principal Investigators: Franceschi
  - UADT Subsites: Oral
  - Control source: Hospital-based
  - Cases: 611
  - Controls: 643
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **Boston (HNSCC)**
  - Boston - US
  - Principal Investigators: Kelsey
  - UADT Subsites: Hospital-based
  - Cases: 513
  - Controls: 593
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **University of Pittsburgh (SCCHN-SPORE)**
  - Pittsburgh - US
  - Genotyping centre: University of Pittsburgh
  - Principal Investigators: Romkes
  - UADT Subsites: Hospital-based
  - Cases: 610
  - Controls: 771
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

- **The Netherlands**
  - Maastricht Hospital - Netherlands
  - Genotyping centre: University St Radboud
  - Principal Investigators: Lacko/Peters
  - UADT Subsites: Hospital-based
  - Cases: 454
  - Controls: 304
  - Includes countries: Poland, Russia, Slovakia, Czech Republic.

#### Total

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,744</td>
<td>11,982</td>
</tr>
</tbody>
</table>

- 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
INHANCE consortium. As previously described [1,3,47], all INHANCE studies have extensive information on tumor site and histology, as well as lifestyle characteristics. The Szczecin, Seattle, UCLA and MD Anderson studies were only able to genotype a proportion of the variants (Table S5). Results for the three ADH variants, rs1229984, rs1573496 and rs698 have been published previously for the Latin American study (LA). For these variants, in “replication” analysis the Latin American study was excluded. All studies have been approved by local ethics committees as well as IARC IRB.

Replication genotyping

Replication genotyping was performed using the TaqMan genotyping platform in 8 participating genotyping laboratories (Table 4). The robustness of the Taqman assays (primers and probes are available upon request) were confirmed at IARC by re-genotyping the CEPH HapMap (CEU) trios and confirming concordance with HapMap genotypes. Any discordance between Hapmap and Taqman generated genotypes was resolved by direct DNA sequencing. All Taqman assays were found to be performing robustly. IARC supplied Taqman assays and a standardized Taqman genotyping protocol to each of the 8 participating genotyping laboratories. A common series of 90 standard DNAs were genotyped at each laboratory to ensure the quality and comparability of the genotyping results across the different studies. Concordance with the consensus genotype and the results produced at the eight genotyping laboratories for the standardized DNAs was 99.7%, and no individual centre had a overall discordance of less than 99.5%. If the assay produced 2 or more discordant genotypes relative to the consensus, the study genotypes for this genetic variant were not included in the statistical analysis. Assays that had a per-centre success rate of <90% or for which genotype distributions deviated from HWE (p<0.001) were also excluded (Table S5).

Replication statistical analysis

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 ADH1/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 (and 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), Banskia 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/cgi/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 Pharoah and SEARCH for contribution of biological samples.
Author Contributions


References