nature genetics

Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer

Corina Lesseur[1], Brenda Diergaarde[2,3], Andrew F Olshan[4,5], Victor Wünsch-Filho[6], Andrew R Ness[7,8], Geoffrey Liu[9], Martin Lacko[10], José Eluf-Neto[11], Silvia Franceschi[1], Pagona Lagiou[12], Gary J Macfarlane[13], Lorenzo Richiardi[14], Stefania Boccia[15], Jerry Polesel[16], Kristina Kjaerheim[17], David Zaridze[18], Mattias Johansson[1], Ana M Menezes[19], Maria Paula Curado[20], Max Robinson[21], Wolfgang Ahrens[22], Cristina Canova[23], Ariana Znaor[1,24], Xavier Castellsagué[25,51], David I Conway[26,27], Ivana Holcátová[28], Dana Mates[29], Marta Vilensky[30], Claire M Healy[31], Neonila Szeszenia-Dąbrowska[32], Eleonóra Fabiánová[33], Jolanta Lissowska[34], Jennifer R Grandis[35,36], Mark C Weissler[37], Eloiza H Tajara[38], Fabio D Nunes[39], Marcos B de Carvalho[40], Steve Thomas[8], Rayjean J Hung[41], Wilbert H M Peters[42], Rolando Herrero[1], Gabriella Cadoni[43], H Bas Bueno-de-Mesquita[44–46], Annika Steffen[47], Antonio Agudo[48], Oxana Shangina[18], Xiangjun Xiao[49], Valérie Gaborieau[1], Amélie Chabrier[1], Devasena Anantharaman[1], Paolo Boffetta[50], Christopher I Amos[49], James D McKay[1] & Paul Brennan[1]

We conducted a genome-wide association study of oral cavity and pharyngeal cancer in 6,034 cases and 6,585 controls from Europe, North America and South America. We detected eight significantly associated loci ($P < 5 \times 10^{-8}$), seven of which are new for these cancer sites. Oral and pharyngeal cancers combined were associated with loci at 6p21.32 (rs3828805, *HLA-DQB1*), 10q26.13 (rs201982221, *LHPP*) and 11p15.4 (rs1453414, *OR52N2–TRIM5*). Oral cancer was associated with two new regions, 2p23.3 (rs6547741, *GPN1*) and 9q34.12 (rs928674, *LAMC3*), and with known cancer-related loci—9p21.3 (rs8181047, *CDKN2B-AS1*) and 5p15.33 (rs10462706, *CLPTM1L*). Oropharyngeal cancer associations were limited to the human leukocyte antigen (HLA) region, and classical HLA allele imputation showed a protective association with the class II haplotype HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 (odds ratio (OR) = 0.59, $P = 2.7 \times 10^{-9}$). Stratified analyses on a subgroup of oropharyngeal cases with information available on human papillomavirus (HPV) status indicated that this association was considerably stronger in HPV-positive (OR = 0.23, $P = 1.6 \times 10^{-6}$) than in HPV-negative (OR = 0.75, $P = 0.16$) cancers.

Cancers of the oral cavity (OC) and oropharynx (OPC) are predominantly caused by tobacco and alcohol use, although oral infection with HPV, particularly HPV16, is an increasingly important cause of OPC[1], especially in the United States and northern Europe[1,2]. The proportion of HPV-related OPC cases varies widely and is estimated to be approximately 60% in the United States, 30% in Europe and lower in South America[2–5]. Genetic factors have also been implicated in OC and OPC susceptibility, especially polymorphisms within alcohol-related genes, including *ADH1B* (alcohol dehydrogenase 1B) and *ADH7* (alcohol dehydrogenase 7)[6,7]. To identify additional susceptibility loci, 13,107 individuals from 12 epidemiological studies (**Supplementary Table 1**) were genotyped using the Illumina OncoArray; after stringent quality control steps (Online Methods and **Supplementary Table 2**), 6,034 cases and 6,585 cancer-free controls remained for analyses (**Table 1**). We next performed genome-wide imputation using the Haplotype Reference Consortium panel[8] and obtained approximately 7 million high-quality imputed variants (**Supplementary Fig. 1**). Given the ancestry diversity of our study, we evaluated associations separately for each continent (Europe, North America and South America) using multivariate unconditional logistic regressions under a log-additive genetic model adjusted for age, sex and regional eigenvectors. Results by continent were combined using fixed-effects meta-analyses to derive associations for overall OC and pharynx cancer (oral, oropharynx, hypopharynx and overlapping cancers; $n = 6,034$), as well as site-specific OC ($n = 2,990$) and OPC ($n = 2,641$). Although several ancestry groups are present in the study, supervised ancestry analyses indicated that >90% of participants were predominantly of European (>70% CEU) ancestry, although some population admixture was observed in South America (**Supplementary Table 3**).

Genome-wide association meta-analyses of overall and site-specific cancers identified nine regions at genome-wide significance ($P < 5 \times 10^{-8}$) (**Fig. 1**). Quantile–quantile plots of observed versus expected P values showed moderate genomic inflation (λ) for the three meta-analyses (λ range = 1.04–1.06; **Supplementary Figs. 2** and **3**). As λ increases with sample size, we scaled it to 1,000 cases and controls, resulting in ameliorated inflation ($\lambda_{1,000}$ range = 1.009–1.01)[9]. Overall OC and pharynx cancer were associated with rs79767424 (5p14.3),

A full list of affiliations appears at the end of the paper.

Received 2 May; accepted 1 September; published online 17 October 2016; doi:10.1038/ng.3685

Table 1 Epidemiological and clinical characteristics of cases and controls

	Cases		Controls	
	n	%	*n*	%
Total	6,034		6,585	
Tumor site				
Oral cavity	2,990	49.6		
Oropharynx	2,641	43.8		
Hypopharynx	305	5.1		
Overlapping	73	1.2		
Other	25	0.4		
Geographic region				
Europe	2,499	41.4	2,928	44.5
North America	2,549	42.2	2,522	38.3
South America	986	16.3	1,135	17.2
Sex				
Male	4,527	75.02	4,325	65.68
Female	1,507	24.98	2,260	34.32
Age, years				
≤50	1,315	21.8	1,355	20.6
50–60	2,006	33.2	1,954	29.7
60–70	1,748	29.0	1,983	30.1
≥70	964	16.0	1,293	19.6
Unknown	1	0.02		
Smoking status				
Never	1,057	17.5	2,508	38.1
Former	1,792	29.7	2,263	34.4
Current	2,623	43.5	1,466	22.3
Unknown	562	9.3	348	5.3
Drinking status				
Never	820	13.6	1,199	18.2
Ever	4,840	80.2	4,840	73.5
Unknown	374	6.2	546	8.3

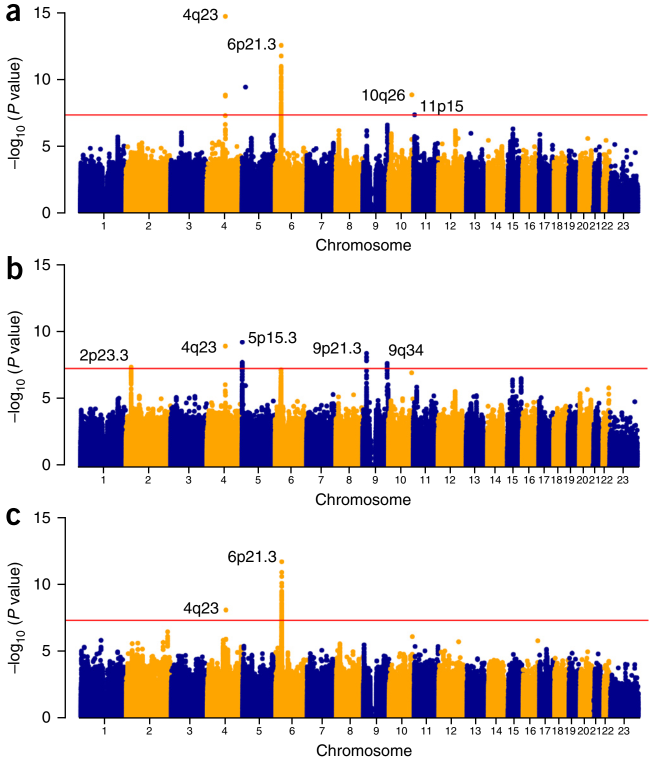

Figure 1 Genome-wide association meta-analysis results. Red lines correspond to $P = 5 \times 10^{-8}$. The *y* axes show $-\log_{10} P$ values. (**a**) Overall OC and pharyngeal cancer analysis with 6,034 cases and 6,585 controls. (**b**) OC analysis with 2,990 cases and 6,585 controls. (**c**) Oropharyngeal cancer analysis with 2,641 cases and 6,585 controls. Loci with technically validated genome-wide significant SNPs are tagged by genomic location.

rs1229984 (4q23), rs201982221 (10q26.13), rs1453414 (11p15.4) and 123 SNPs at 6p21.32 (**Supplementary Table 4**). Twenty-six variants were associated ($P < 5 \times 10^{-8}$) with OC (**Supplementary Table 5**), with 4 mapping to 2p23.3, 1 mapping to 4q23, 3 mapping to 9q34.12, 13 mapping to 5p15.33 and 5 mapping to 9p21.3. For OPC, new significant variants were located at 6p21.32 (62 SNPs; **Supplementary Table 6**). Suggestive susceptibility variants ($P < 5 \times 10^{-7}$; **Supplementary Tables 7–9**) were associated with OC at four additional loci—6p21.33, 6p21.32, 15q21.2 and 15q26.2—and with OPC at 2q36.1. Other genomic locations outside the HLA region showed promising associations ($P < 5 \times 10^{-6}$) with OPC (**Supplementary Table 10**). For susceptibility loci at $P < 5 \times 10^{-8}$, functional annotations of regulatory features with Encyclopedia of DNA Elements (ENCODE) and expression quantitative trait locus (eQTL) information, if available,

Table 2 Lead genome-wide significant SNP for each validated locus from regional meta-analyses of oral and pharyngeal cancers combined and analysis of each cancer separately

Region	SNP	Chr:pos[a]	Gene	EA/OA[b]	INFO (R^2)[c]	AF[d] cases/ controls	OR	*P*	P_{het}
Oral cancer and pharyngeal cancer									
4q23	rs1229984	4:100239319	*ADH1B*	A/G	G	0.03/0.06	0.56	2.29×10^{-15}	0.002
6p21.32	rs3828805	6:32636120	*HLA-DQB1*	C/T	0.88	0.75/0.72	1.28	3.35×10^{-13}	0.007
10q26.13	rs201982221	10:126157446	*LHPP*	D/I	G	0.03/0.02	1.67	1.58×10^{-9}	0.50
11p15.4	rs1453414	11:5829084	*OR52N2–TRIM5*	C/A	G	0.23/0.20	1.19	4.78×10^{-8}	0.55
Oral cancer									
2p23.3	rs6547741	2:27855924	*GPN1*	A/G	0.98	0.50/0.54	0.83	3.97×10^{-8}	0.34
4q23	rs1229984	4:100239319	*ADH1B*	A/G	G	0.03/0.06	0.57	1.09×10^{-9}	0.02
5p15.33	rs10462706	5:1343794	*CLPTM1L*	T/C	0.97	0.12/0.15	0.74	5.54×10^{-10}	0.84
9p21.3	rs8181047	9:22064465	*CDKN2B-AS1*	A/G	G	0.29/0.24	1.24	3.80×10^{-9}	0.37
9q34.12	rs928674	9:133952024	*LAMC3*	G/A	0.89	0.14/0.12	1.33	2.09×10^{-8}	0.88
Oropharyngeal cancer									
4q23	rs1229984	4:100239319	*ADH1B*	A/G	G	0.02/0.06	0.55	8.53×10^{-9}	0.05
6p21.32	rs3828805	6:32636120	*HLA-DQB1*	C/T	0.88	0.75/0.72	1.37	2.21×10^{-12}	0.07

[a]SNP position according to NCBI Genome Build 37 (hg19). [b]EA, effect allele; OA, other allele. [c]G, genotyped; SNP INFO (R^2) is the average across imputation batches. [d]AF, allele frequency of the effect allele.

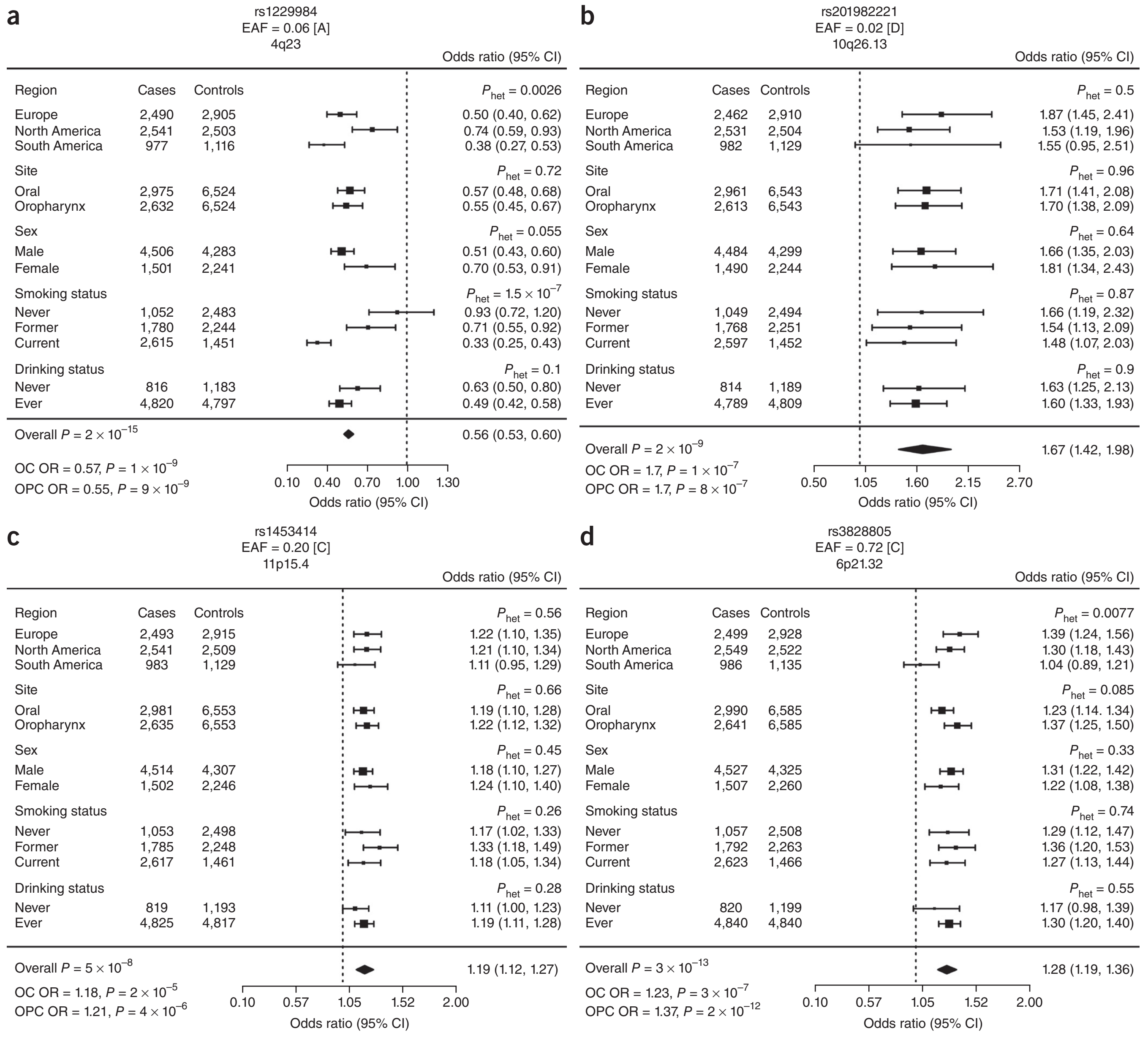

Figure 2 Forest plots of odds ratios for the lead SNP at each genome-wide significant locus in the overall oral and pharyngeal cancer meta-analysis. (**a**–**d**) Results are shown for the loci at 4q23, rs1229984 (**a**), 10q26.13, rs20198222 (**b**), 11p15.4, rs1453414 (**c**) and 6p21.32, rs3828805 (**d**). CI, confidence interval; EAF, effect allele frequency in 6,585 controls. The effect allele appears in brackets.

are summarized in **Supplementary Tables 11** and **12**. Given the geographic heterogeneity of our population, we performed sensitivity analyses after excluding individuals with <70% CEU ancestry; these analyses gave similar results (**Supplementary Table 13**). To validate array genotypes and imputed dosages, we directly genotyped at least one variant within each locus ($P = 5 \times 10^{-7}$) in a subset of approximately 700 individuals using a different platform (TaqMan). The concordance between genotyped or imputed genotypes and TaqMan results was >97% for all regions, with the exception of rs2398180, an imputed variant that had concordance of 94% (**Supplementary Tables 14** and **15**). TaqMan assays could not be designed for two rare variants, rs201982221 (10q26.13) and rs7976742 (5p14.3), and we used Sanger sequencing for validation (**Supplementary Table 16**). We were able to validate the rs201982221 deletion (Online Methods), but rs7976742 did not validate (Online Methods). The lead variants at each validated locus ($P < 5 \times 10^{-8}$) for overall and site-specific analyses are shown in **Table 2**. Results stratified by geographical region, smoking and alcohol status are displayed in **Figure 2** (oral and pharynx cancer combined) and **Figure 3** (oral cancer).

The association at rs1229984 (4q23, *ADH1B*) has previously been reported as a susceptibility locus for OC and OPC[6]; as in previous findings, this variant showed heterogeneity by region, smoking status and alcohol drinking status[10,11] (**Fig. 2a**). Three other 4q23 SNPs reached $P < 5 \times 10^{-8}$, although conditional analyses indicated that these are not independent signals (**Supplementary Table 17**). The rs1573496 (*ADH7*) variant reported to be strongly associated with OC and OPC in the previous genome-wide association study (GWAS) for upper aerodigestive tract cancer[7] was

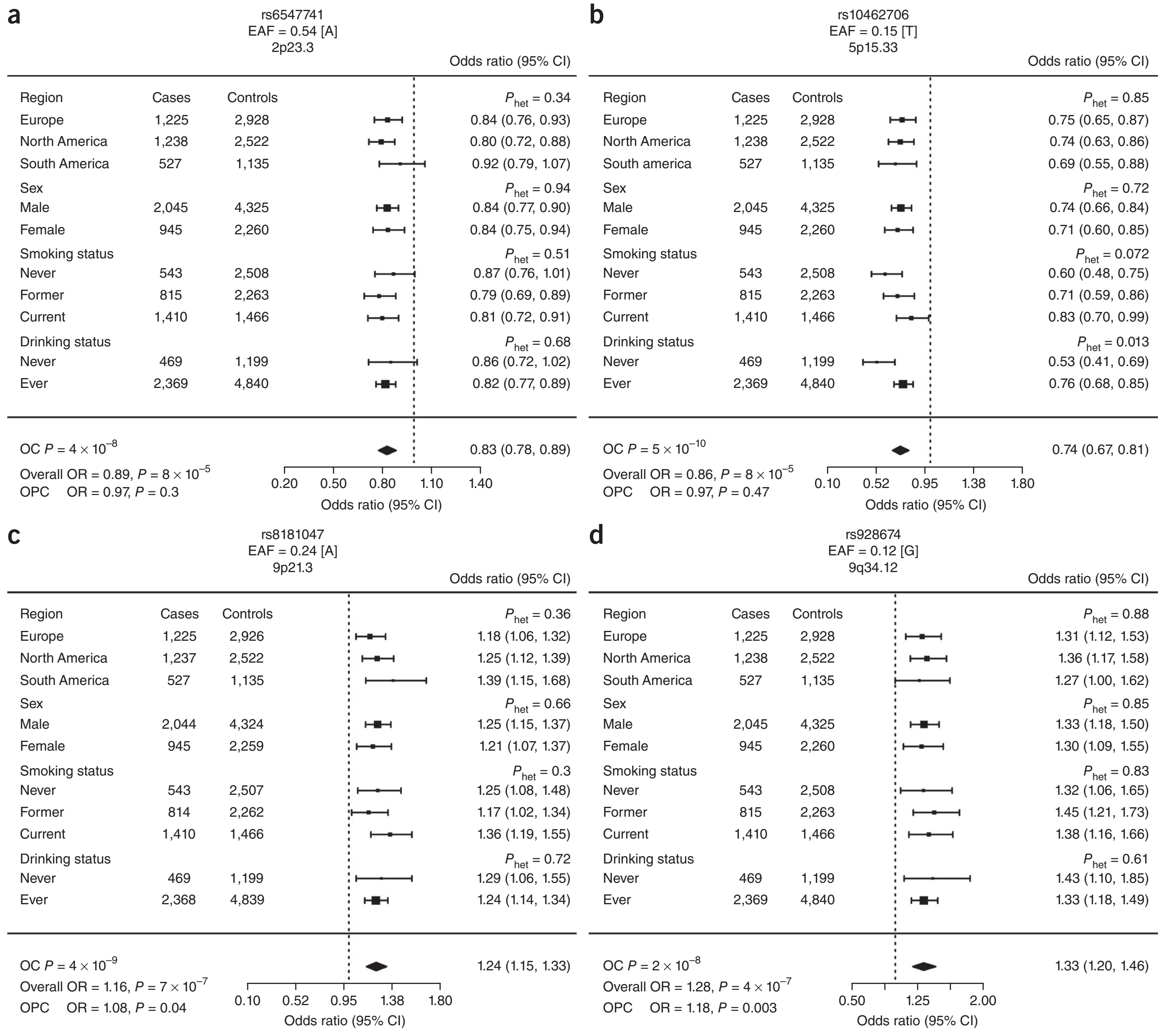

Figure 3 Forest plots of odds ratios for the lead SNP at each genome-wide significant locus in the oral cancer meta-analysis. (**a**–**d**) Results are shown for 2p23.3, rs6547741 (**a**), 5p15.33, rs10462706 (**b**), 9p21.3, rs8181047 (**c**) and 9q34.12, rs928674 (**d**). EAF, effect allele frequency in 6,585 controls. The effect allele appears in square brackets. "Overall" denotes results for oral and pharyngeal cancer.

only moderately associated here (**Supplementary Table 18**). In the overall OC and pharynx cancer analysis, we identified rs201982221 at 10q26.13 (OR = 1.67, $P = 1.58 \times 10^{-9}$), which was also separately associated with OC (OR = 1.71, $P = 1.04 \times 10^{-7}$) and OPC (OR = 1.70, $P = 7.9 \times 10^{-7}$) (**Fig. 2b**). rs201982221 is located in the *LHPP* gene in a region with reported regulatory features (**Supplementary Table 11**). However, it is a rare intronic deletion in an area of low linkage disequilibrium (LD) (**Supplementary Fig. 4**) and thus warrants further validation in a different population. rs1453414, the lead signal at 11p15.4 (**Supplementary Table 19**), is an intronic variant that showed borderline association in the overall (OR = 1.19, $P = 4.78 \times 10^{-8}$) and site-specific (OC (OR = 1.19, $P = 1.65 \times 10^{-5}$) and OPC (OR = 1.22, $P = 4.26 \times 10^{-6}$)) analyses (**Fig. 2c** and **Supplementary Fig. 5**). rs1453414 is upstream of *OR52N2*, which encodes an olfactory receptor, and within *TRIM5*, which encodes an E3 ubiquitin ligase, and is an eQTL for these genes in brain tissue[12] (**Supplementary Table 12**).

At 2p23.3, four SNPs showed evidence ($P < 5 \times 10^{-8}$) of association with OC and in conditional analyses did not appear to be independent (**Supplementary Table 20**). These signals map to an area with high LD that includes *C2orf16*, *ZNF512*, *CCDC121* and *GPN1* (**Supplementary Fig. 6**). The lead SNP, rs6547741, was associated with OC but not with OPC and maps to an intron of *GPN1*, which encodes a GTPase involved in RNA polymerase II transport and DNA repair[13]. Associations between rs6547741 and OC were homogenous across other analyses stratified by region, sex, smoking status and drinking status (**Fig. 3a**).

Variation at 5p15.33 was also exclusively associated with OC susceptibility (OPC: rs10462706, $P = 0.47$). The top signal, rs10462706, was associated with decreased OC risk (OR = 0.74, $P = 5.54 \times 10^{-10}$) and is

Table 3 Association of the HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 haplotype in individuals with >70% European ancestry

	Haplotype[a] cases/controls	Cases/controls	HF[b] cases	HF[b] controls	OR	P	Meta-analysis[c]	
							OR	P
Oral and pharynx cancer							0.68	3.32×10^{-10}
Europe	207/422	2,497/2,928	0.04	0.07	0.60	4.04×10^{-8}		
North America	207/276	2,342/2,329	0.04	0.06	0.74	1.68×10^{-3}		
South America	74/101	613/727	0.06	0.07	0.86	0.35		
Oral cancer							0.75	1.72×10^{-4}
Europe	106/422	1,231/2,928	0.04	0.07	0.60	1.52×10^{-5}		
North America	128/276	1,135/2,329	0.06	0.06	0.92	4.67×10^{-1}		
South America	41/101	351/727	0.06	0.07	0.80	0.26		
Oropharynx cancer							0.59	2.73×10^{-9}
Europe	84/422	1,098/2,928	0.04	0.07	0.57	2.69×10^{-5}		
North America	72/276	1,119/2,329	0.03	0.06	0.52	3.49×10^{-6}		
South America	31/101	216/727	0.07	0.07	1.05	0.81		
Oropharynx cancer by HPV status[d]								
HPV positive	11/505	336/3,686	0.01	0.07	0.23	1.6×10^{-6}		
HPV negative	25/505	240/3,686	0.05	0.07	0.75	0.16		

[a]Number of copies of the haplotype in cases and controls. [b]Haplotype frequency (HF) calculated as the total number of copies of the haplotype in the population (haplotype copies/2n). [c]Fixed-effects meta-analysis of regional associations adjusted for age, sex and eigenvectors. [d]HPV status available in a subset of cases from the ARCAGE, EPIC, CHANCE and Pittsburgh studies.

in low LD ($r^2 = 0.15$; **Supplementary Fig. 7**) with the second strongest signal, rs467095. These two variants are 7 kb apart and map to intron 13 of *CLPTM1L*. In stratified analysis, they showed stronger effects in never smokers (heterogeneity P value (P_{het}) = 0.07 and 0.0028, respectively) and never drinkers (P_{het} = 0.01 and 0.0025, respectively) (**Fig. 3b** and **Supplementary Fig. 8**). Conditional analyses showed that these SNPs were not completely independent (**Supplementary Table 21**). *TERT* and *CLPTM1L* encode telomerase reverse transcriptase (TERT) and the cleft lip and palate–associated transmembrane 1–like protein (CLPTM1L), respectively. Notably, rs467095 is an esophageal eQTL for *TERT*[14] (**Supplementary Table 12**) and is in high LD with rs401681 (OR = 1.18, $P = 2.1 \times 10^{-7}$, $r^2 = 0.94$), a widely studied SNP associated with risk of several cancers, including: lung[15,16], bladder, prostate, cervical, melanoma[17], basal cell[18], esophageal[19], pancreatic[20] and nasopharyngeal[21] cancers. Multiple 5p15.33 variants have been reported to independently influence cancer risk, both increasing and decreasing risk. Interestingly, rs401681[A] was associated with increased OC risk, similar to previous melanoma associations and in the opposite direction to previous lung cancer results[17].

Several variants within the *CDKN2A*–*CDKN2B* locus (9p21.3) were found to be associated with OC. The lead SNP, rs8181047, is an intronic variant within *CDKN2B-AS1* (*CDKN2B* antisense RNA 1) (**Fig. 3c**). rs8181047 is in LD (range r^2 = 0.6–0.8) with four other 9p21.3 variants strongly associated with OC (**Supplementary Fig. 9**) that in conditional analyses did not show independent associations (**Supplementary Table 22**). The *CDKN2A*–*CDKN2B* locus contains genes involved in cell cycle regulation and senescence and has been associated with multiple malignancies, including melanoma[22], glioma[23], basal cell[18], breast[24], lung[25], nasopharyngeal[26] and esophageal[27] cancers. Notably, *CDKN2A* is frequently mutated in HPV-negative head and neck cancers[28].

The OC-associated variants at 9q34.12 mapped to an intron of *LAMC3*, which encodes a laminin involved in cortical development[29]. rs928674, the peak signal, showed consistent effects across strata and a weaker association with OPC ($P = 0.003$) (**Fig. 3d**). rs928674 is in high LD with three other robustly associated 9q34.12 SNPs (range r^2 = 0.82–0.96; **Supplementary Fig. 10** and **Supplementary Table 23**) and is an esophageal mucosa *cis*-eQTL for a downstream gene, *AIF1L* (allograft inflammatory factor 1 like)[14].

The most prominent finding in the overall and OPC meta-analyses was a strong association signal at 6p21.32 within the HLA class II region. The lead variant in both analyses, rs3828805, maps 1.7 kb 5′ of *HLA-DQB1* (**Fig. 2d** and **Supplementary Fig. 11**) and, similar to other 6p21.32 variants (**Supplementary Tables 4** and **6**), showed heterogeneity in effect by geographic region (P_{het} = 0.007) with no effect in South America ($P = 0.62$). Association analyses of 6p21.32 variants ($P < 5 \times 10^{-8}$) conditioned on rs3828805 did not show multiple independent signals (**Supplementary Table 24**), suggesting a common haplotype. To further investigate HLA associations, we imputed classical alleles in 11,436 individuals (with >70% European ancestry) (Online Methods). Three classical HLA alleles, HLA-DRB1*1301, HLA-DQA1*0103 and HLA-DQB1*0603, reached $P < 5 \times 10^{-8}$ in the overall analysis and were also strongly associated with OPC (**Supplementary Table 25**). These alleles are in high LD (r^2 >0.9) and are part of the HLA class II haplotype HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603, which is common in Europeans and was previously reported to be associated with decreased cervical cancer risk[30]. The HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 haplotype was strongly associated with reduced OPC risk (OR = 0.59, $P = 2.7 \times 10^{-9}$) and was more weakly associated with OC risk (OR = 0.75, $P = 1.7 \times 10^{-4}$). Further conditional analysis on this haplotype and 6p21.32 variants did not show evidence of additional independent effects (**Supplementary Table 26**). Given the importance of HPV infection in the etiology of cervical and oropharyngeal cancers[31], we conducted *post-hoc* analyses to examine the effect of HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 haplotype in a subset of 576 cases with information available on HPV status and 3,662 controls. HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 was associated with strongly reduced risk of HPV-positive OPC (OR = 0.23, $P = 1 \times 10^{-6}$, $n = 336$), with no significant association in 240 HPV-negative OPC cases (OR = 0.75, $P = 0.16$) (**Table 3**). These results indicate that the class II HLA region is implicated in at least two HPV-driven cancers, namely HPV-positive OPC and cervical cancer. The lack of an association between 6p21.3 SNPs and OPC risk in South America could relate to previous findings that less than 10% of OPC cases are positive for HPV in this region[4,5]. Moreover, weaker association with OC could be due to a smaller proportion of the cases being positive for HPV, as well as to some OPC cases possibly being misclassified, especially for tumors at the base of the tongue. Further evaluation of the extent and specificity of this HLA effect in HPV-associated cancers is important given the strength of the observed association. Such evaluation may help elucidate why some individuals are at higher risk of HPV-positive OPC after HPV infection and may also have implications for cancer immunotherapies targeting the HLA class II antigen presentation pathway[32].

In summary, we identified seven oral and pharyngeal cancer susceptibility loci, including a strong HLA signal narrowed to a class II haplotype. Future replication of these findings in an independent population is warranted, as well as fine-mapping and functional studies necessary to establish the biological framework underlying these associations.

URLs. R software, http://www.r-project.org/; PLINK, https://www.cog-genomics.org/plink2; University of Michigan Imputation Server, https://imputationserver.sph.umich.edu/; Haplotype Reference Consortium, http://www.haplotype-reference-consortium.org/; SNP2HLA, https://www.broadinstitute.org/mpg/snp2hla/; HaploReg v4.1, http://www.broadinstitute.org/mammals/haploreg/haploreg.php; GTEx portal, http://www.gtexportal.org/home/; LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/; metafor R package, http://www.metafor-project.org/doku.php; LDlink, http://analysistools.nci.nih.gov/LDlink/; INHANCE Consortium, http://www.inhance.utah.edu/; OncoArray Network, http://epi.grants.cancer.gov/oncoarray/; GAME-ON, http://epi.grants.cancer.gov/gameon/; GENCAPO, http://www.gencapo.famerp.br.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Genotype data for the oral and pharyngeal OncoArray study have been deposited at the database of Genotypes and Phenotypes (dbGaP) under accession phs001202.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Genotyping performed at the Center for Inherited Disease Research (CIDR) was funded through US National Institute of Dental and Craniofacial Research (NIDCR) grant 1X01HG007780-0. Genotyping for shared controls with the Lung OncoArray initiative was funded through grant X01HG007492-0. C.L. undertook this work during the tenure of a postdoctoral fellowship awarded by the International Agency for Research on Cancer. The funders did not participate in study design, data collection and analysis, decision to publish or preparation of the manuscript. We acknowledge all of the participants involved in this research and the funders and support. We thank L. Fernandez for her contribution to the IARC ORC multicenter study. We are also grateful to S. Koifman for his contribution to the IARC Latin America multicenter study (S. Koifman passed away in May 2014) and to X. Castellsagué who recently passed away (June 2016).

The University of Pittsburgh head and neck cancer case–control study is supported by US National Institutes of Health grants P50CA097190 and P30CA047904. The Carolina Head and Neck Cancer Study (CHANCE) was supported by the National Cancer Institute (R01CA90731). The Head and Neck Genome Project (GENCAPO) was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; grants 04/12054-9 and 10/51168-0). The authors thank all the members of the GENCAPO team. The HN5000 study was funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10034); the views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the UK Department of Health. The Toronto study was funded by the Canadian Cancer Society Research Institute (020214) and the National Cancer Institute (U19CA148127) and by the Cancer Care Ontario Research Chair. The Alcohol-Related Cancers and Genetic Susceptibility Study in Europe (ARCAGE) was funded by the European Commission's fifth framework programme (QLK1-2001-00182), the Italian Association for Cancer Research, Compagnia di San Paolo/FIRMS, Region Piemonte and Padova University (CPDA057222). The Rome Study was supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC) awards IG 2011 10491 and IG 2013 14220 to S.B. and by Fondazione Veronesi to S.B. The IARC Latin American study was funded by the European Commission INCO-DC programme (IC18-CT97-0222), with additional funding from Fondo para la Investigación Científica y Tecnológica (Argentina) and the Fundação de Amparo à Pesquisa do Estado de São Paulo (01/01768-2). The IARC Central Europe study was supported by the European Commission's INCO-COPERNICUS Program (IC15-CT98-0332), US NIH/National Cancer Institute grant CA92039 and World Cancer Research Foundation grant WCRF 99A28.The IARC Oral Cancer Multicenter study was funded by grant S06 96 202489 05F02 from Europe against Cancer; grants FIS 97/0024, FIS 97/0662 and BAE 01/5013 from Fondo de Investigaciones Sanitarias, Spain; the UICC Yamagiwa-Yoshida Memorial International Cancer Study; the National Cancer Institute of Canada; Associazione Italiana per la Ricerca sul Cancro; and the Pan-American Health Organization. Coordination of the EPIC study is financially supported by the European Commission (DG SANCO) and the International Agency for Research on Cancer.

AUTHOR CONTRIBUTIONS

P. Brennan and J.D.M. conceived and designed the project. C.L. undertook data harmonization, genotypes quality control, GWAS analysis, imputation and meta-analyses. X.X. performed genotype calling. V.G. and A.C. organized and supervised sample selection and DNA shipments at the International Agency for Research on Cancer. A.C. performed replication TaqMan genotyping. C.L. and V.G. analyzed data from replication genotyping. C.L. and P. Brennan drafted the first version of the manuscript. B.D., A.F.O., V.W.-F., A.R.N., G.L., M.L., J.E.-N., S.F., P.L., G.J.M., L.R., S.B., J.P., K.K., D.Z., M.J., A.M.M., M.P.C., M.R., W.A., C.C., A.Z., X.C., D.I.C., I.H., D.M., M.V., C.M.H., N.S.-D., E.F., J.L., J.R.G., M.C.W., E.H.T., F.D.N., M.B.d.C., S.T., R.J.H., W.H.M.P., R.H., G.C., A.S., A.A., O.S., H.B.B.-d.-M., P. Boffetta and D.A. contributed with reagents, samples and/or materials and reviewed and approved the final manuscript. J.D.M. and C.I.A. designed and coordinated the Lung Cancer OncoArray. P. Brennan obtained funding for the project and provided overall supervision and management.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gillison, M.L., Chaturvedi, A.K., Anderson, W.F. & Fakhry, C. Epidemiology of human papillomavirus–positive head and neck squamous cell carcinoma. *J. Clin. Oncol.* **33**, 3235–3242 (2015).
2. Chaturvedi, A.K. *et al.* Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J. Clin. Oncol.* **29**, 4294–4301 (2011).
3. Kreimer, A.R., Clifford, G.M., Boyle, P. & Franceschi, S. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review. *Cancer Epidemiol. Biomarkers Prev.* **14**, 467–475 (2005).
4. Ribeiro, K.B. *et al.* Low human papillomavirus prevalence in head and neck cancer: results from two large case–control studies in high-incidence regions. *Int. J. Epidemiol.* **40**, 489–502 (2011).
5. López, R.V. *et al.* Human papillomavirus (HPV) 16 and the prognosis of head and neck cancer in a geographical region with a low prevalence of HPV infection. *Cancer Causes Control* **25**, 461–471 (2014).
6. Hashibe, M. *et al.* Multiple ADH genes are associated with upper aerodigestive cancers. *Nat. Genet.* **40**, 707–709 (2008).
7. McKay, J.D. *et al.* A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet.* **7**, e1001333 (2011).
8. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
9. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**R2, R122–R128 (2008).
10. Hashibe, M. *et al.* Evidence for an important role of alcohol- and aldehyde-metabolizing genes in cancers of the upper aerodigestive tract. *Cancer Epidemiol. Biomarkers Prev.* **15**, 696–703 (2006).
11. Chang, J.S., Straif, K. & Guha, N. The role of alcohol dehydrogenase genes in head and neck cancers: a systematic review and meta-analysis of *ADH1B* and *ADH1C*. *Mutagenesis* **27**, 275–286 (2012).
12. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
13. Carré, C. & Shiekhattar, R. Human GTPases associate with RNA polymerase II to mediate its nuclear import. *Mol. Cell. Biol.* **31**, 3953–3962 (2011).
14. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
15. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406 (2008).
16. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407–1409 (2008).
17. Rafnar, T. *et al.* Sequence variants at the *TERT–CLPTM1L* locus associate with many cancer types. *Nat. Genet.* **41**, 221–227 (2009).

18. Stacey, S.N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat. Genet.* **41**, 909–914 (2009).
19. Yin, J. *et al. TERT–CLPTM1L* rs401681 C>T polymorphism was associated with a decreased risk of esophageal cancer in a Chinese population. *PLoS One* **9**, e100667 (2014).
20. Petersen, G.M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* **42**, 224–228 (2010).
21. Bei, J.X. *et al.* A GWAS meta-analysis and replication study identifies a novel locus within *CLPTM1L/TERT* associated with nasopharyngeal carcinoma in individuals of Chinese ancestry. *Cancer Epidemiol. Biomarkers Prev.* **25**, 188–192 (2016).
22. Law, M.H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat. Genet.* **47**, 987–995 (2015).
23. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899–904 (2009).
24. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
25. Timofeeva, M.N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum. Mol. Genet.* **21**, 4980–4995 (2012).
26. Bei, J.X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet.* **42**, 599–603 (2010).
27. Wu, C. *et al.* Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nat. Genet.* **46**, 1001–1006 (2014).
28. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
29. Barak, T. *et al.* Recessive *LAMC3* mutations cause malformations of occipital cortical development. *Nat. Genet.* **43**, 590–594 (2011).
30. Chen, D. *et al.* Genome-wide association study of susceptibility loci for cervical cancer. *J. Natl. Cancer Inst.* **105**, 624–633 (2013).
31. Bouvard, V. *et al.* A review of human carcinogens—Part B: biological agents. *Lancet Oncol.* **10**, 321–322 (2009).
32. Thibodeau, J., Bourgeois-Daigneault, M.C. & Lapointe, R. Targeting the MHC class II antigen presentation pathway in cancer immunotherapy. *OncoImmunology* **1**, 908–916 (2012).

[1]International Agency for Research on Cancer (IARC/WHO), Lyon, France. [2]Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [3]University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania, USA. [4]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [5]UNC Lineberger Comprehensive Cancer Center, Chapel Hill, North Carolina, USA. [6]Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brazil. [7]National Institute for Health Research (NIHR) Biomedical Research Unit in Nutrition, Diet and Lifestyle at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, UK. [8]School of Oral and Dental Sciences, University of Bristol, Bristol, UK. [9]Princess Margaret Cancer Centre, Toronto, Ontario, Canada. [10]Department of Otorhinolaryngology, Head and Neck Surgery, Maastricht University Medical Center, Maastricht, the Netherlands. [11]Departamento de Medicina Preventiva, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [12]Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece. [13]School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK. [14]Department of Medical Sciences, University of Turin, Turin, Italy. [15]Section of Hygiene, Institute of Public Health, Università Cattolica del Sacro Cuore, Fondazione Policlinico 'Agostino Gemelli', Rome, Italy. [16]Unit of Cancer Epidemiology, CRO Aviano National Cancer Institute, Aviano, Italy. [17]Cancer Registry of Norway, Oslo, Norway. [18]Department of Cancer Epidemiology and Prevention, Institute of Carcinogenesis, N.N. Blokhin Russian Cancer Research Centre of the Russian Ministry of Health, Moscow, Russian Federation. [19]Programa de Pós-Graduacão em Epidemiologia, Universidade Federal de Pelotas (UFPel), Pelotas, Brazil. [20]Epidemiology, International Center for Research (CIPE), A.C. Camargo Cancer Center, Sao Paulo, Brazil. [21]Centre for Oral Health Research, Newcastle University, Newcastle, UK. [22]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany. [23]Deparment of Molecular Medicine, University of Padova, Padova, Italy. [24]Croatian National Institute of Public Health, Zagreb, Croatia. [25]Institut Català d'Oncologia (ICO)–IDIBELL, CIBERESP, l'Hospitalet de Llobregat, Barcelona, Spain. [26]School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, UK. [27]NHS National Services Scotland (NSS), Edinburgh, UK. [28]Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic. [29]National Institute of Public Health, Bucharest, Romania. [30]Instituto de Oncología 'Angel H. Roffo', Universidad de Buenos Aires, Buenos Aires, Argentina. [31]Trinity College School of Dental Science, Dublin, Ireland. [32]Department of Environmental Epidemiology, Nofer Institute of Occupational Medicine, Lodz, Poland. [33]Regional Authority of Public Health, Banská Bystrica, Slovakia. [34]Maria Skłodowska-Curie Memorial Cancer Centre and Institute of Oncology (MCMCC), Warsaw, Poland. [35]Department of Otolaryngology–Head and Neck Surgery, University of California at San Francisco, San Francisco, California, USA. [36]Clinical Translational Science Institute, University of California at San Francisco, San Francisco, California, USA. [37]Department of Otolaryngology/Head and Neck Surgery, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [38]Department of Molecular Biology, School of Medicine of São José do Rio Preto, São José do Rio Preto, Brazil. [39]Department of Stomatology, School of Dentistry, University of São Paulo, São Paulo, Brazil. [40]Department of Head and Neck Surgery, Heliópolis Hospital, São Paulo, Brazil. [41]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [42]Department of Gastroenterology, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands. [43]Institute of Otorhinolaryngology, Università Cattolica del Sacro Cuore, Fondazione Policlinico 'Agostino Gemelli', Rome, Italy. [44]Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands. [45]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [46]Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia. [47]German Institute of Human Nutrition in Potsdam-Rehbruecke (DIfE), Nuthetal, Germany. [48]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Institut Català d'Oncologia (ICO)–IDIBELL, l'Hospitalet de Llobregat, Barcelona, Spain. [49]Department of Biomedical Data Sciences, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA. [50]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [51]Deceased. Correspondence should be addressed to P. Brennan (gep@iarc.fr).

ONLINE METHODS

Study population and genotyping. The study population comprised 6,034 cases and 6,585 controls derived from 12 epidemiological studies, the majority of case–control design and part of the International Head and Neck Cancer Epidemiology Consortium (INHANCE). Cases and controls from a European cohort study (EPIC) and cases from a UK case series (HN5000) were also included. The characteristics and references for each study are summarized in **Supplementary Table 1**. Informed consent was obtained for all participants, and studies were approved by respective institutional review boards. Cancer cases comprised the following ICD codes: oral cavity (C02.0–C02.9, C03.0–C03.9, C04.0–C04.9, C05.0–C06.9), oropharynx (C01.9, C02.4, C09.0–C10.9), hypopharynx (C13.0–C13.9) and overlapping (C14 and combination of the codes for other sites); 25 oral or pharyngeal cases had unknown ICD codes (other).

Genomic DNA isolated from blood or buccal cells was genotyped at the Center for Inherited Disease Research (CIDR) with the Illumina OncoArray custom designed for cancer studies by the OncoArray Consortium[33], part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. The majority of the samples were genotyped as part of the oral and pharynx cancer OncoArray, with the exception of 2,476 shared controls (1,453 from the EPIC study and 1,023 from the Toronto study) that were genotyped at CIDR but as part of the Lung OncoArray. Genotype calls were made by the Dartmouth team in GenomeStudio software (Illumina) using a standardized cluster file for OncoArray studies. Cluster plots for top SNPs for individuals genotyped as part of the oral and pharynx cancer OncoArray are shown in **Supplementary Figure 12**.

GWAS quality control. We used PLINK 1.934 to conduct systematic quality control steps on genotypes calls. An initial filtering step on the complete data set excluded samples with genotyping rates <80% and SNPs with call rates <80%. We then excluded samples and SNPs with call rates <95%. During quality control for individuals, we removed samples with unsolved discrepancies between genetic and reported sex and individuals with an outlier autosomal heterozygosity rate (±4 s.d. from the mean). Identity-by-descent (IBD) analysis performed on the LD-pruned data set (r^2 <0.1) identified 103 expected experimental duplicate pairs (IBD >0.9). For each of these pairs, we excluded the sample with the lower genotyping rate, and 74 unexpected duplicate pairs were excluded. Additionally, we identified 44 unexpected pairs of relatives (IBD >0.3) and excluded one sample from each pair, prioritizing cases over controls; for pairs with the same status, we excluded one sample according to genotyping rate. After the initial SNP filtering by call rate and exclusion of duplicated and zeroed probes, 513,311 probes remained for analysis. Next, given the heterogeneity of the study population, we divided the data set by geographic region (Europe, North America and South America) and excluded SNPs within each region with deviation from Hardy–Weinberg equilibrium in controls ($P < 1 \times 10^{-7}$). To account for potential population stratification within each geographic area, we performed principal-components analysis (PCA) in EIGENSTRAT[34] using approximately 10,000 common markers in low LD (r^2 <0.004, minor allele frequency (MAF) >0.05); subsequently, we excluded population outliers (n = 139) and derived regional eigenvectors to adjust regional GWAS analyses (**Supplementary Fig. 13**). PCA plots for each individual study are displayed in **Supplementary Figure 14**. To examine ancestry within the regions, we used STRUCTURE 2.3.4 (ref. 35) under a supervised approach with HapMap samples to determine the CEU, YRI, and CHB + JPT ancestry for each individual. All coordinates refer to genome build hg19/GRCh37.

Imputation. Imputation of unknown genetic variation was performed using the Michigan Imputation Server[36], a free service for large-scale population studies. We used SHAPEIT[37] for prephasing, Minimac3 (ref. 38) for imputation and the first release of the Haplotype Reference Consortium panel[8], a large collection of human haplotypes (n = 64,976) that combines data from multiple initiatives, including the 1000 Genomes Project. For imputation, we used a set of high-quality SNPs with MAF >0.01 and call rate >98% that we imputed in randomized batches of approximately 3,000 individuals and analyzed imputation quality statistics together. SNPs with MAF <0.01 or R^2 <0.3 in any of the batches were excluded before association analysis. Thus, the final set of 7,099,472 imputed variants used in association analysis was of very high quality: 86% of the variants with MAF ≥ 0.05 had R^2 >0.9, and 52% of the less common variants (MAF <0.05) had R^2 >0.9 (**Supplementary Fig. 1**).

GWAS and meta-analyses. We undertook an approach involving GWAS by geographic region and meta-analysis to evaluate the relationship between SNPs and overall OC and pharynx cancer, as well as site-specific OC and OPC risk; in total, we tested 7,574,753 SNPs (genotyped and imputed variants). GWAS analysis in each region was performed in PLINK with the R glm function for genotype dosages, using multivariable unconditional logistic regression assuming a log-additive genetic or dosage model with age, sex and principal components as covariates. Next, association statistics were included in a fixed-effects meta-analysis performed in PLINK[39]. P values for heterogeneity were calculated using Cochran's Q. Conditional analyses within associated regions were performed in R (glm), and meta-analysis of regional results was performed in PLINK.

HLA imputation and haplotype analyses. Using SNP2HLA[40] and the Type I Diabetes Genetics Consortium reference panel of 5,225 individuals of European descent, we imputed HLA classical alleles and amino acids into 11,436 study participants with >70% CEU ancestry. Individuals were randomized into 12 groups of approximately 1,000 individuals each, and imputation quality scores from Beagle were examined across groups. Markers with $R^2 \geq 0.7$ in all groups and average R^2 >0.9 (8,286 of the 8,961 markers included in the panel) were carried on for analyses, and 98% of these markers had R^2 >0.95. We then performed regional association analyses of binary markers followed by meta-analysis of imputed binary markers (SNPs, classical alleles and amino acids) as previously described using PLINK and R. HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 haplotype information was extracted from Beagle files of phased best-guess genotypes and used to estimate odds ratios and 95% confidence intervals from multivariable unconditional logistic regression models for each phenotype. Haplotype- and HPV-stratified analyses, in a subset of cases with available data, were performed with unconditional logistic regression where haplotype was the predictor and HPV-positive or HPV-negative OPC was the outcome. The measure of HPV positivity was derived from HPV16E6 serology in the ARCAGE and EPIC studies and from tumor HPV16 DNA and p16 overexpression in the CHANCE study; for the Pittsburgh cases, a combination of tumor HPV *in situ* hybridization and p16 immunohistochemistry was used.

Technical validation genotyping. To confirm array genotypes and imputed dosages, we genotyped at least one SNP within each associated ($P < 5 \times 10^{-7}$) genomic region using TaqMan assays (Thermo Fisher Scientific) in a subset of approximately 900 individuals from the ARCAGE, IARC Latin America, EPIC and IARC OC studies (**Supplementary Table 14**). TaqMan genotyping included standard positive and negative controls; assay numbers and cycling conditions are described in **Supplementary Table 15**. For 10 of the 12 SNPs genotyped by TaqMan assay, there was no evident departure from Hardy–Weinberg equilibrium ($P > 0.05$). However, two variants, rs467095 and rs3731239, were slightly out of equilibrium only in samples from the EPIC study ($P < 0.01$), which might relate to the small sample size genotyped in the technical validation. The rare variants at 5p14.3 (rs79767424) and 10q26.13 (rs201982221; **Supplementary Fig. 15**) could not be genotyped with TaqMan assays; thus, for these variants, validation was performed with Sanger sequencing on a small subset of samples available within the IARC studies: rs201982221 was assessed in 15 wild-type, 15 heterozygous and 2 homozygous individuals, and rs79767424 was assessed in 15 wild-type and 15 heterozygous individuals. Primer sequences and conditions for PCR are described in **Supplementary Table 16**; amplification products were sequenced on the Applied Biosystems 3730 DNA Analyzer at Biofidal, France.

Bioinformatic annotation and analyses. To explore the possible functional implications of variants within associated regions, we used an internet tool, HaploReg v4.1 (ref. 41), that includes annotations from ENCODE. Functional annotation for variants reaching $P < 5 \times 10^{-8}$ is summarized in **Supplementary Table 11**. We also retrieved eQTL information from several sources, including the GTEx Project[14] and HaploReg[41]. Manhattan and quantile–quantile plots were generated in R using the package qqman[42], and forest plots were generated using the metafor R library[43]. Regional association plots were generated using LocusZoom[44] with LD and recombination rate data from the 1000 Genomes Project November 2014 release. To search for additional information on pairwise LD, we used the online tool LD link[45].

33. Anonymous. Consortium launches genotyping effort. *Cancer Discov.* **3**, 1321–1322 (2013).
34. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
35. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
36. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
37. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
38. Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
39. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
40. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
41. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
42. Turner, S.D. qqman: an R package for visualizing GWAS results using Q–Q and Manhattan plots. Preprint at *bioRxiv* http://dx.doi.org/10.1101/005165 (2014).
43. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 48 (2010).
44. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
45. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

doi:10.1038/ng.3685