

1 **Cross-ancestral GWAS identifies 29 novel variants across Head and Neck Cancer**  
2 **subsites**

3  
4 Ebrahimi E<sup>1,2</sup>, Sangphukieo A<sup>1,3</sup>, Park HA<sup>1</sup>, Gaborieau V<sup>1</sup>, Ferreiro-Iglesias A<sup>1</sup>, Diergaard B<sup>4</sup>,  
5 Ahrens W<sup>5</sup>, Alemany L<sup>6,7,8</sup>, Arantes LMRB<sup>9</sup>, Betka J<sup>10</sup>, Bratman SV<sup>11</sup>, Canova C<sup>12</sup>, Conlon  
6 MSC<sup>13</sup>, Conway DI<sup>14</sup>, Cuello M<sup>15</sup>, Curado M<sup>16</sup>, de Carvalho A<sup>1</sup>, de Oliveira J<sup>17</sup>, Gormley M<sup>18</sup>,  
7 Hadji M<sup>2,19</sup>, Hargreaves S<sup>20</sup>, Healy CM<sup>21</sup>, Holcatova I<sup>22</sup>, Hung RJ<sup>23,24</sup>, Kowalski LP<sup>25,26</sup>, Laggiou  
8 P<sup>27</sup>, Laggiou A<sup>28</sup>, Liu G<sup>29</sup>, Macfarlane GJ<sup>30</sup>, Olshan AF<sup>31</sup>, Perdomo S<sup>1</sup>, Pinto LF<sup>32</sup>, Podesta JV<sup>33</sup>,  
9 Polesel J<sup>34</sup>, Pring M<sup>18</sup>, Rashidian H<sup>2</sup>, Gama RR<sup>35</sup>, Richiardi L<sup>36</sup>, Robinson M<sup>37</sup>, Rodriguez-  
10 Urrego PA<sup>38</sup>, Santi SA<sup>39</sup>, Saunders DP<sup>40</sup>, Soares-Lima SC<sup>41</sup>, Timpson N<sup>42</sup>, Vilensky M<sup>43</sup>, von  
11 Zeidler SV<sup>44</sup>, Waterboer T<sup>45</sup>, Zendejdel K<sup>2</sup>, Znaor A<sup>46</sup>, Brennan P<sup>1</sup>, HEADSpAcE  
12 Consortium<sup>47</sup>, McKay J<sup>1</sup>, Virani S<sup>1\*</sup>, Dudding T<sup>18\*</sup>.

13  
14 **\*Co-last, co-corresponding authors**

15 \*Shama Virani  
16 [viranis@iarc.who.int](mailto:viranis@iarc.who.int)

17  
18 \*Tom Dudding

19 [tom.dudding@bristol.ac.uk](mailto:tom.dudding@bristol.ac.uk)

20 <sup>1</sup>Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France,  
21 <sup>2</sup>Cancer Research Center, Cancer Institute, Tehran University of Medical Sciences, Tehran, Iran, <sup>3</sup>Center of  
22 Multidisciplinary Technology for Advanced Medicine (CMUTEAM), Faculty of Medicine, Chiang Mai University,  
23 Chiang Mai, Thailand, <sup>4</sup>Department of Human Genetics, School of Public Health, University of Pittsburgh, and  
24 UPMC Hillman Cancer Center, Pittsburgh, USA, <sup>5</sup>Leibniz Institute for Prevention Research and Epidemiology-  
25 BIPS, Bremen, Germany, <sup>6</sup>Catalan Institute of Oncology. ICO, L'Hospitalet, Barcelona, Spain, <sup>7</sup>Bellvitge Biomedical  
26 Research Institute (IDIBELL), L'Hospitalet, Barcelona, Spain, <sup>8</sup>CIBER en Epidemiología y Salud Pública  
27 (CIBERESP), Madrid, Spain, <sup>9</sup>Barretos Cancer Hospital, Barretos, Brazil, <sup>10</sup>Department of Otorhinolaryngology and  
28 Head And Neck Surgery, 1.st Medical Faculty, Charles University, Faculty Hospital Motol, Prague, Czech Republic,  
29 <sup>11</sup>Department of Radiation Oncology, Princess Margaret Cancer Centre, University of Toronto, Toronto, Canada,  
30 <sup>12</sup>Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thoraco-Vascular Sciences and  
31 Public Health, University of Padova, Padova, Italy, <sup>13</sup>Epidemiology, Outcomes & Evaluation Research, Health  
32 Sciences North Research Institute, Sudbury, Canada, <sup>14</sup>School of Medicine, Dentistry and Nursing, University of  
33 Glasgow, Glasgow, UK, <sup>15</sup>Oncology, Hospital de Clinicas Dr. Manuel Quintela, Montevideo, Uruguay,  
34 <sup>16</sup>Epidemiology and Statistics Group, Research Center, A.C Camargo Cancer Center, São Paulo, Brazil, <sup>17</sup>Araújo  
35 Jorge Cancer Hospital, Associação de Combate ao Câncer em Goiás, Goiania, Brazil, <sup>18</sup>Bristol Dental School,  
36 Bristol University, Bristol, UK, <sup>19</sup>A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio,  
37 Finland, <sup>20</sup>University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK, <sup>21</sup>School of Dental Science,  
38 Dublin Dental University Hospital, Trinity College Dublin, Dublin, Ireland, <sup>22</sup>Institute of Hygiene & Epidemiology, 1st  
39 Faculty of Medicine, Charles University, Prague, Czech Republic, <sup>23</sup>Prosserman Centre for Population Health  
40 Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada, <sup>24</sup>Division of  
41 Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, <sup>25</sup>Department of Head  
42 and Neck Surgery, University of São Paulo Medical School, São Paulo, Brazil, <sup>26</sup>Department of Head and Neck

43 Surgery and Otorhinolaryngology, A C Camargo Cancer Center, São Paulo, Brazil, <sup>27</sup>Department of Hygiene,  
44 Epidemiology & Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens,  
45 Greece, <sup>28</sup>Department of Public and Community Health, School of Public Health, University of West Attica, Athens,  
46 Greece, <sup>29</sup>Medicine, Epidemiology, Medical Oncology, Princess Margaret Cancer Centre, University of Toronto,  
47 Toronto, Canada, <sup>30</sup>Epidemiology Group, School of Medicine, Medical Sciences and Nutrition, University of  
48 Aberdeen, Aberdeen, UK, <sup>31</sup>Department of Epidemiology, Gillings School of Global Public Health, University of  
49 North Carolina, Chapel Hill, USA, <sup>32</sup>Programa de Carcinogênese Molecular, Instituto Nacional de Câncer - INCA,  
50 Rio de Janeiro, Brazil, <sup>33</sup>Head and Neck Surgery Division, Women's Association for Education and Fight Against  
51 Cancer/AFECC, Vitória, Brazil, <sup>34</sup>Unit of Cancer Epidemiology, Centro di Riferimento Oncologico di Aviano (CRO)  
52 IRCCS, Aviano, Italy, <sup>35</sup>Department of Head and Neck Surgery, Barretos Cancer Hospital, São Paulo, Brazil,  
53 <sup>36</sup>Cancer Epidemiology Unit, University of Turin, Turin, Italy, <sup>37</sup>Cellular Pathology, The Newcastle upon Tyne  
54 Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK, <sup>38</sup>Pathology and Laboratories, Pathology, University  
55 Hospital Fundacion Santa Fe de Bogota, Bogota, Colombia, <sup>39</sup>Clinical Oncology Research, Health Sciences North  
56 Research Institute, Sudbury, Canada, <sup>40</sup>Department of Dental Oncology, Health Sciences North, Northern Ontario  
57 School of Medicine University, Sudbury, Canada, <sup>41</sup>Brazilian National Cancer Institute, Rio de Janeiro, Brazil,  
58 <sup>42</sup>MRC Integrative Epidemiology Unit, Bristol University, Bristol, UK, <sup>43</sup>Instituto de Oncologia Angel H Roffo,  
59 Universidad de Buenos Aires, Buenos Aires, Argentina, <sup>44</sup>Pathology Department, Federal University of Espírito  
60 Santo, Vitória, Brazil, <sup>45</sup>Infections and Cancer Epidemiology, German Cancer Research Center (Deutsches  
61 Krebsforschungszentrum, DKFZ), Heidelberg, Germany, <sup>46</sup>Cancer Surveillance Branch, International Agency for  
62 Research on Cancer (IARC/WHO), Lyon, France, <sup>47</sup>A list of authors and their affiliations appears at the end of the  
63 paper.

64

## 65 **Acknowledgements and funding:**

66 This study was funded in part by the European Union's Horizon 2020 research and innovation  
67 program under grant agreement No 825771 (HEADSpAcE project) and by the US National  
68 Institute of Dental and Craniofacial Research (NIDCR) grants R03DE030257 and  
69 R01DE025712. Genotyping using the Oncoarray and the All of Us array was performed at  
70 Center for Inherited Disease (CIDR) and funded by NIDCR 1X01HG007780-0 and jointly by  
71 NIDCR/NCI X01HG010743.

72 This publication presents data from the Head and Neck 5000 study. The study was a  
73 component of independent research funded by the National Institute for Health and Care  
74 Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-  
75 10034). The views expressed in this publication are those of the author(s) and not necessarily  
76 those of the NHS, the NIHR or the Department of Health. Core funding was also provided  
77 through awards from Above and Beyond, University Hospitals Bristol and Weston Research  
78 Capability Funding and the NIHR Senior Investigator award to Professor Andy Ness. Round  
79 1 genotyping was funded by US National Institute of Dental and Craniofacial Research  
80 (NIDCR) grant 1X01HG007780-0. Round 2 genotyping was funded by World Cancer

81 Research Fund Pilot Grant (grant number: 2018/1792), Above and Beyond Charity (GA2500),  
82 Wellcome Trust Research Training Fellowship (201237/Z/16/Z) and Cancer Research UK  
83 Programme Grant, the Integrative Cancer Epidemiology Programme (grant number:  
84 C18281/A19169). This latter grant also supported Human papillomavirus (HPV) serology. This  
85 research has been conducted using the UK Biobank Resource under Application Number  
86 40644. The work of Dr. Polesel is partially supported by Italian Ministry of Health 'Ricerca  
87 Corrente'.

88 The University of Pittsburgh head and neck cancer case-control study is supported by US  
89 National Institutes of Health grants P50CA097190, P30CA047904 and R01DE025712.

90 Geoffrey Liu is the M. Qasim Choksi Research Chair in Translational Research at University  
91 Health Network and University of Toronto and is supported by the Princess Margaret Head  
92 and Neck Translational Program, which is supported by philanthropic funds from the Wharton  
93 Family, Joe's Team, Gordon Tozer, Reed Fund, and the Riley Family.

94 The University of North Carolina studies were supported in part by grants CA61188 and  
95 CA90731 from the National Institutes of Health.

96 Northern Cancer Foundation (Principal Investigator Grants to MSC Conlon, DP Saunders).

97 Rayjean J. Hung is the CIHR Canada Research Chair and the study is supported by the  
98 Canadian Cancer Society and Canadian Institute of Health Research.

99 The authors would like to thank all the patients, and their families involved in these studies.

100 Where members are identified as personnel of the International Agency for Research on  
101 Cancer/ World Health Organization, the authors alone are responsible for the views expressed  
102 in this article and they do not necessarily represent the decisions, policy or views of the  
103 International Agency for Research on Cancer / World Health Organization.

104 **Conflict of Interest Disclosure:**

105 Tim Waterboer serves on advisory boards for MSD (Merck) Sharp & Dohme.

106 Scott V Bratman is inventor on patents related to cell-free DNA mutation and methylation  
107 analysis technologies that are unrelated to this work and have been licensed to Roche and  
108 Adela, respectively. Scott V Bratman is a co-founder of and has ownership in Adela.

109 **Abstract (192 words)**

110 In this multi-ancestry genome-wide association study (GWAS) and fine mapping study of head  
111 and neck squamous cell carcinoma (HNSCC) subsites, we analysed 19,073 cases and 38,857  
112 controls and identified 29 independent novel loci. We provide robust evidence that a 3' UTR  
113 variant in *TP53* (rs78378222, T>G) confers a 40% reduction in odds of developing overall  
114 HNSCC. We further examine the gene-environment relationship of *BRCA2* and *ADH1B*  
115 variants demonstrating their effects act through both smoking and alcohol use. Through  
116 analyses focused on the human leukocyte antigen (HLA) region, we highlight that although  
117 human papilloma virus (HPV)(+) oropharyngeal cancer (OPC), HPV(-) OPC and oral cavity  
118 cancer (OC) all show GWAS signal at 6p21, each subsite has distinct associations at the  
119 variant, amino acid, and 4-digit allele level. We also defined the specific amino acid changes  
120 underlying the well-known DRB1\*13:01-DQA1\*01:03-DQB1\*06:03 protective haplotype for  
121 HPV(+) OPC. We show greater heritability of HPV(+) OPC compared to other subsites, likely  
122 to be explained by HLA effects. These findings advance our understanding of the genetic  
123 architecture of head and neck squamous cell carcinoma, providing important insights into the  
124 role of genetic variation across ancestries, tumor subsites, and gene-environment interactions.

125 **Main**

126 Head and neck squamous cell carcinomas (HNSCC) are a heterogeneous group of cancers  
127 originating primarily in the oral cavity (OC), oropharynx (OPC), larynx (LA) and hypopharynx  
128 (HPC). Currently, HNSCC is ranked the 6th most common cancer globally, although incidence  
129 is predicted to increase 30% by 2030<sup>1,2</sup>. Tobacco smoking and alcohol consumption are major  
130 risk factors, particularly in high income countries, contributing to 72% of cases when used  
131 together, while betel quid/areca nut products significantly increase risk in some Asia-Pacific  
132 populations<sup>2</sup>. HNSCC subsites can be differentially affected, with smoking more strongly  
133 linked to laryngeal cancer and drinking more strongly linked to OC/OPC<sup>3</sup>. There has been a  
134 decline in smoking in high income countries, as such, the increasing incidence could be due  
135 to changes in etiology<sup>4</sup>. Infection with human papillomavirus (HPV), particularly HPV type 16,  
136 is a recently identified causal risk factor for OPC<sup>5-7</sup> and the proportion of HPV-associated  
137 OPCs is highest in high-income countries (63%-85%)<sup>8</sup>. Disparities in epidemiology, risk, and  
138 prognosis highlight the recognition of HPV-associated OPC as a distinct biological entity<sup>9</sup>.

139 Although a limited number of genome-wide association studies (GWAS) have been conducted  
140 on HNSCC, a germline contribution to HNSCC risk has been established, with multiple  
141 susceptibility loci associated with risk. These include the 4q23 locus (*ADH1B*, *ADH7*) linked  
142 to genes involved in alcohol metabolism, the 5p15 locus (*TERT-CLPTM1L*) associated with

143 genes responsible for DNA stability maintenance, and the 6p21 and 6p22 loci, mostly within  
144 the human leukocyte antigen (HLA) region, corresponding with genes regulating the innate  
145 immune response<sup>10–14</sup>. The 6p locus within the HLA region has been a specific area of focus  
146 for HPV-driven cancers, with the hypothesis being that variants influencing immune response  
147 to viral antigens would be most relevant for risk<sup>10,13,14</sup>. However, there is an emerging role for  
148 the immune microenvironment for other HNSCC subsites<sup>15</sup>, suggesting that the HLA may  
149 confer risk separately in other HNSCC subsites, potentially via non-HPV mechanisms.  
150 Previous GWAS were limited in sample size for HNSCC subsites making inference between  
151 subsites, particularly for HLA, difficult. They were also conducted predominantly in subjects of  
152 European ancestry, limiting generalizability of findings.

153 Despite knowledge of the major risk factors and several risk loci for HNSCC, identifying those  
154 who will develop cancer is still difficult. Not all smokers develop cancer and risk loci only offer  
155 a fractional change in risk at the population level. The interaction between environmental  
156 factors and risk loci may help explain additional risk and have been reported for lung cancer  
157 (smoking)<sup>16</sup>, colorectal cancer (alcohol)<sup>17</sup> and bladder cancer (arsenic exposure)<sup>18</sup> among  
158 others. Studies investigating these interactions need large sample sizes and individual level  
159 exposure data harmonised across studies which, is often not possible in large GWAS meta-  
160 analyses.

161 Here, we perform a cross-ancestry GWAS of HNSCC using individual level data, bringing  
162 together studies from Europe, North America, South America, South Asia and the Middle East.  
163 We identify multiple novel genetic risk susceptibility loci, determine shared and unique risk loci  
164 across subsites, explore interactions between genetic and environmental factors in HNSCC  
165 risk and conduct fine mapping of the HLA region. This work lays the foundation for identifying  
166 HNSCC susceptibility loci with increased representation from non-European populations.

## 167 **Results**

### 168 **Cross-ancestral meta-analysis identifies 18 novel genetic loci across HNSCC subsites.**

169 In this cross-ancestral meta-analysis of two pooled individual level datasets (Table S1), we  
170 evaluated 13,092,551 genetic variants in 19,073 HNSCC cases and 38,857 controls. Of the  
171 HNSCC cases, there were 5,596 (29%) oral cavity (OC), 5,411 (28%) oropharyngeal (OPC),  
172 4,409 (23%) laryngeal (LA), 898 (5%) hypopharyngeal (HPC), 2,759 (14%) unknown (either  
173 unknown primary site or not available) or overlapping sites. HPV status was available for 68%  
174 of OPC cases, of which 3,685 (60%) were HPV(+) (Table S2).

175 We identified 18 novel genome-wide associated variants, including two specific to non-  
176 European ancestry (Table 1, Figure 1, Figure S1) and validated 6 previously identified loci  
177 (Table S3, S4). A specific focus on the HLA region identified 11 further novel variants. The  
178 proportion of variance attributable to genome-wide SNPs for HNSCC overall was 14% (95%  
179 Confidence Interval (CI): 12.7, 15.3). Across subsites, heritability ranged from 7.6% (95% CI:  
180 5.0, 10.2) for HPC to 29.1% (95% CI: 25.5, 32.7) for HPV(+) OPC (Table S5).

181 For overall HNSCC, two novel variants in the 1q32 region were identified. rs61817953, near  
182 *PIK3C2B*, was associated with decreased risk (OR (95%CI)=0.90 (0.87, 0.93),  $p_{\text{meta}}=2.17 \times 10^{-8}$ )  
183 <sup>8</sup>) and rs6679311 near *MDM4*, a strong negative regulator of p53, was associated with  
184 increased risk (OR (95% CI)=1.11 (1.07, 1.14),  $p_{\text{meta}}=1.25 \times 10^{-10}$ ) (Figure S2). The latter is in  
185 moderately high LD ( $r^2=0.75$ ) with rs4245739, an *MDM4* 3' UTR variant known to increase  
186 breast<sup>19</sup> and prostate<sup>20</sup> cancer risk. At the 13q13 locus, rs7334543, a novel 3' UTR variant in  
187 *BRCA2* was associated with decreased risk of overall HNSCC (OR (95%CI)= 0.91 (0.88,  
188 0.94),  $p_{\text{meta}}=2.39 \times 10^{-8}$ ) and was independent from rs11571833, stop gain variant previously  
189 identified in this region for UADTs<sup>12</sup>. Within those of European ancestry, rs78378222 a 3' UTR  
190 variant in *TP53*, was associated with a reduced risk of HNSCC overall, (OR (95% CI)=0.62  
191 (0.52, 0.73),  $p=2.16 \times 10^{-8}$ ) (Figure 2a). The effect was consistent across all non-HPV related  
192 HNSCC subtypes but had no effect in HPV(+) OPC. The T>G allele frequency of rs78378222  
193 is 0.01 in EUR, 0.002 in AFR and AMR populations, and nearly absent in all other 1000  
194 Genomes super-populations; as such, there was no effect of this variant in the Mixed ancestry.  
195 Given its low frequency, technical validation was performed in 2,370 samples and  
196 concordance with imputed data was 99.9% (Table S6). There was strong evidence for this  
197 variant being correlated with decreased gene expression of *TP53* (Figure 2b) (Table S7). This  
198 variant is in the poly-adenylation signal of the *TP53* gene and potentially leads to impaired 3'  
199 end processing of *TP53* mRNA<sup>21</sup>. rs78378222 is located within a highly conserved sequence  
200 (TTTTATTGTAAAATA -> TTGTATTGTAAAATA) that appears to be crucial for miRNA binding.  
201 This region is predicted to interact with 5 different microRNAs (miRNAs) as suggested by  
202 TarBase (<https://dianalab.e-ce.uth.gr/tarbasev9>) (Figure 2c).

203 For OC, three novel loci were identified (Table 1). First, rs28419191, an intergenic variant at  
204 5q31 associated with an increased risk of OC (OR (95% CI)=1.23 (1.15, 1.31),  $p_{\text{meta}}=3.16 \times 10^{-10}$ )  
205 <sup>10</sup>). This variant was in high LD with rs1131769 ( $r^2=0.93$ ), a missense variant in *STING1* which  
206 was a novel loci for overall HNSCC risk (OR=1.13 (1.09, 1.18),  $p_{\text{meta}}=2.38 \times 10^{-10}$ ) (Table 1,  
207 Figure 2d). Both rs28419191 and rs1131769 correlated with expression of catenin alpha 1  
208 (*CTNNA1*), a gene related to RNA and actin filament binding, but not *STING1* expression in  
209 whole blood; as such, the function of this variant is unclear (Figure 2e). The second novel

210 variant rs67351073, located at 20q13 in Zinc Finger CCCH-Type And G-Patch Domain  
211 Containing (*ZGPAT*), was associated with reduced risk of OC (OR (95%CI)=0.78 (0.72, 0.85),  
212  $p_{\text{meta}}=4.45 \times 10^{-8}$ ). A highly correlated variant seen within the European ancestry only  
213 (rs4809325,  $r^2=0.97$ ), which also decreased OC risk, was correlated with decreased *ZGPAT*  
214 gene expression in whole blood (PP4 score=0.97) and increased *LIME1* gene expression in  
215 oesophageal and lung mucosa (Figure S3, Table S7). Finally, a novel, rare, European  
216 ancestry-specific intronic variant, rs577454702, located in the mitogen-activated protein  
217 kinase 1 (*MAPK1*) gene at 22q11, was associated with a large increased risk of OC (OR  
218 (95%CI)=2.60 (1.86, 3.65),  $p=2.53 \times 10^{-8}$ ).

219 For laryngeal cancer, rs55831773, a novel splice variant, mapping to *ATP1B2* was associated  
220 with increased risk (OR (95% CI)=1.21 (1.13, 1.29),  $p_{\text{meta}}=5.1 \times 10^{-9}$ ). *ATP1B2* is in close  
221 proximity to *TP53* but conditional analyses suggest this variant is independent of the rare *TP53*  
222 3' UTR variant described for overall HNSCC. There was also no evidence that rs55831773  
223 alters *TP53* expression, further suggesting independent effects of these two variants (Figure  
224 S4). A novel intronic variant, rs10419397, located in a gene dense region of 19p13 was also  
225 strongly associated with LA (OR (95%CI)=1.13 (1.10, 1.17),  $p_{\text{meta}}=1.21 \times 10^{-14}$ ). This variant has  
226 been found to associate with mitochondrial dysfunction<sup>22,23</sup> and is in very high LD with several  
227 variants associated with risk of other cancers, including rs4808616 ( $r^2>0.99$ ), a 3' UTR for  
228 *ABHD8* linked to breast and lung cancers<sup>24</sup>. rs200410709 is a novel variant which showed  
229 increased risks in the Mixed ancestry but with no evidence of effect in Europeans. It is a  
230 deletion variant in an intergenic region, adjacent to the Syntaxin Binding Protein 6 (*STXBP6*)  
231 gene (14q12), and was associated with increased risk of LA (OR (95% CI)=3.38 (2.26, 5.07),  
232  $p=3.57 \times 10^{-9}$ ) (Figure S5).

233 Five HPC specific variants were identified for the first time. rs138707495, a rare (MAF:  
234 European = 0.009, Mixed = 0.005) variant located in the 3' UTR of *GDF7* (OR (95%CI)=3.06  
235 (2.07, 4.53),  $p_{\text{meta}}=2.33 \times 10^{-8}$ ), rs77750788 at 11q25 near *IGSF9B* (OR (95%CI)=2.07 (1.61,  
236 2.68),  $p_{\text{meta}}=2.03 \times 10^{-8}$ ) and rs181194133 an intronic variant in *OPCML* (OR (95%CI)= 3.44  
237 (2.24, 5.31),  $p_{\text{meta}}= 2.09 \times 10^{-8}$ ) were all associated with increased risk of HPC in the cross-  
238 ancestral meta-analysis. Within the European group, rs181777026 (11q14) located near  
239 *TENM4* was associated with increased risk of HPC. Conversely, rs150899739 (6q24), which  
240 showed an increased risk in the Mixed ancestry but no effect in Europeans, is within *SASH1*  
241 and greatly increased the risk for HPC (OR (95% CI)=5.84 (3.17, 10.76),  $p=1.47 \times 10^{-8}$ ) (Figure  
242 S6).

243 At 3p21, rs1520483, a novel intronic variant in the lactotransferrin (*LTF*) gene, associated with  
244 an increased risk of HPV(+) OPC (OR (95% CI)=1.23 (1.14, 1.32),  $p=2.19 \times 10^{-8}$ ) in Europeans.  
245 *LTF* acts as a transcription factor, inducing expression of innate immune related genes for  
246 antiviral host defence<sup>25,26</sup>.

247 rs112726671, a variant near the vitamin D receptor (*VDR*) gene, was associated with risk of  
248 HPV(-) OPC (OR (95%CI)=1.23 (1.14, 1.32),  $p_{\text{meta}}=2.19 \times 10^{-8}$ ). This variant is independent  
249 from rs35189640, a nearby variant previously identified to increase risk of HPV(-) OPC  
250 ( $r^2=0.0005$ )<sup>10</sup>.

### 251 Refining previously identified HNSCC risk variants

252 Loci identified in previous GWAS of HNSCCs at 4q23 (*ADH1B*, *ADH1C*, *ADH7*), 5p15  
253 (*CLPTM1L*), 6p21 (HLA), 6p22 (*ZNRD1-AS1*), 9p21 (*CDKN2B-AS1*), 12q24 (*ALDH2*) and  
254 15q21 (*FGF7*) were validated here. Notably, rs11571833 (13q13), the rare (MAF:  
255 Europeans=0.009, Mixed=0.007) stop gained variant, resulting in a stop codon 93 amino acids  
256 early in the *BRCA2* protein, was strongly associated with an increased risk of LA (OR (95%  
257 CI)=2.09 (1.65, 2.66),  $p_{\text{meta}}=1.57 \times 10^{-9}$ ) and HPC (lead variant for HPC- rs11571815: OR (95%  
258 CI)=2.73 (1.61, 3.90),  $p_{\text{meta}}=3.99 \times 10^{-8}$ ) separately. Previous GWAS combining lung and  
259 aerodigestive tract cancers as well as studies using targeted genotyping have found this  
260 variant to substantially increase risk for smoking related cancers<sup>27</sup>, however this is the first  
261 time this variant was identified in within specific subsites.

### 262 Distinct interactions of smoking and alcohol use with risk variants

263 rs11571833 the *BRCA2* stop-gained variant validated here showed clear evidence of a dose-  
264 response effect across smoking and drinking strata, but the variant did not correlate with  
265 variants related to smoking-related behaviours such as smoking initiation or cigarettes per day  
266 in colocalization analysis (Table S7). However, the variant effect was present in both non-  
267 drinking smokers and non-smoking drinkers, suggesting the risk increasing effect of  
268 rs11571833 requires either carcinogenic influence. This *BRCA2* variant shows a similar gene-  
269 environment interaction separately within the European and Mixed ancestries, despite  
270 differences in sample size (Figure 3a).

271 We confirm that rs1229984, the well-described missense variant in the *ADH1B* gene, has a  
272 strong protective effect on OC which is only seen in smokers or in drinkers when stratifying by  
273 use (Figure S6b). However, we measure a strong correlation between rs1229984 and variants  
274 associated with alcoholic drinks per week but not cigarettes per day or smoking initiation (Table



275 S7). To separate out the linked behaviors of smoking and drinking we investigated the  
276 association in combinations of drinking and smoking status. These analyses confirm  
277 rs1229984 has an effect in those who smoke and drink and in non-drinking smokers but not  
278 non-smoking drinkers, suggesting the mechanism through smoking as may be more important  
279 (Figure 3b). Interactions with smoking and drinking for *ADH1C* and *ADH7* were less clear.

280 rs58365910 near *CHRNA5*, known to alter smoking intensity<sup>36</sup> showed a suggestive  
281 association with LA consistent effects across the European and Mixed ancestries (Figure S8).  
282 The increasing risk effect of this variant was correlated with increased smoking intensity and  
283 when evaluated by exposure group, this variant shows a clear interaction with smoking but  
284 not alcohol use (Figure 3c).

### 285 **Novel Loci in the HLA Region specific to oral cavity and oropharynx cancer**

286 Our genome-wide results highlight heterogeneity in the Human Leukocyte Antigen (HLA)  
287 region, which encodes genes involved in immune response, across HNSCC subsites. For  
288 HPV(+) OPC, signals were identified at both 6p21 and 6p22 but for OC only the 6p21 signal  
289 was seen. The HLA region is particularly susceptible to genetic diversity across populations  
290 and is highly polymorphic with a dense LD structure. To account for this, genotyped variants  
291 in this region were re-imputed to define variants, amino acid changes and 4-digit alleles, which  
292 were then analysed separately using fine mapping strategies to identify independent signals.  
293 Independence of signals was carefully evaluated using linkage and conditional analysis  
294 (Tables S8, S9).

295 Overall, 19 independent signals reached significance (Tables S10, S11). Eleven novel risk  
296 variants were identified specific to OC, HPV(+) OPC, HPV(-) OPC, and for HNSCC overall  
297 (Table 2, Figure S9).

298 Three novel intronic variants were associated with risk of HNSCC overall. The Chr6:33046667  
299 variant, near *HLA-DPB1* (OR (95% CI)= 1.11 (1.07, 1.14),  $p_{meta}=1.32 \times 10^{-8}$ ) and rs28360051  
300 near *PSORS1C3* (OR (95% CI)= 1.23 (1.14, 1.34),  $p_{meta}=1.91 \times 10^{-7}$ ) both increased HNSCC  
301 risk. The rs28360051 variant was strongly driven by its effect in HPV(+) OPC. A novel intronic  
302 variant, rs1536036, mapping to *ITPR3*, a receptor that mediates the release of intracellular  
303 calcium, was protective for HNSCC overall (OR (95% CI)= 0.85 (0.80, 0.91),  $p= 8.42 \times 10^{-7}$ )  
304 only in the admixed ancestry.

305 For HPV(+) OPC, five novel variants were identified. rs4143334, in the noncoding transcript  
306 exon of *ZDHHC20P2* increased cancer risk (OR (95% CI)= 1.89 (1.51, 2.35),  $p_{meta}=1.91 \times 10^{-7}$ )

307 <sup>8</sup>). The remaining three had important functional significance. The first (DRB1 37Asn/Ser)  
308 causes an amino acid change in the antigen binding pocket (P9 pocket) of the beta chain of  
309 the HLA-DR protein and reduces HPV(+) OPC risk (OR (95% CI)= 0.68 (0.63, 0.73),  $p_{\text{meta}}=$   
310  $3.22 \times 10^{-23}$ ). The second (HLA-B 67Cys/Ser/Tyr) is in an antigen binding pocket (B-pocket) of  
311 HLA-B and also results in decreased HPV(+) OPC risk (OR (95% CI)=0.81 (0.74, 0.88),  $p_{\text{meta}}=$   
312  $1.33 \times 10^{-6}$ ) (Figure 4a). The third (DRB1 233Thr), is in exon 5 of *DRB1* and increases risk of  
313 HPV(+) OPC (OR (95% CI)=1.27 (1.17, 1.38),  $p_{\text{meta}}= 7.15 \times 10^{-9}$ ). This amino acid change is in  
314 high LD with several others that are in the HLA-DR binding pocket, of which 5 have similar  
315 risk (Table S12). Accuracy of best-fit models, which included each amino acid in place of  
316 DRB1 233Thr, were found to be similar to the original model containing DRB1 233Thr ( $\Delta\text{BIC}$   
317  $\pm 2$ ), indicating that presence of any of these five amino acid changes—including DRB1  
318 10Glu/Gln and 12Lys located in the HLA binding pocket—confers similar levels of risk (Figure  
319 4b, Table S12). Within those of European ancestry, the novel HLA-B\*51:01 allele increased  
320 risk of HPV(+) OPC (OR (95% CI)=1.9 (1.55, 2.31),  $p_{\text{meta}}= 3.6 \times 10^{-10}$ ).

321 For HPV(-) OPC, rs1131212 was found to be associated with an increased risk (OR  
322 (95%CI)=1.33 (1.19, 1.49),  $p_{\text{meta}}=5.33 \times 10^{-7}$ ) (Figure 5a). This novel, functional variant maps to  
323 exon 2 of the *HLA-B* gene causing an amino acid change Gln94His in an *HLA-B* binding  
324 pocket. rs1131212 tags another functional HLA-B amino acid change, HLA-B 70Asn/Ser  
325 (Table S12) in strong LD ( $r^2=1$ ) which has a similar effect and with similar model accuracy (OR  
326 (95%CI)=1.32 (1.18, 1.47),  $p_{\text{meta}}=8.81 \times 10^{-7}$ ) (Figure 5b). These results suggest that presence  
327 of either rs1131212 or HLA-B 70Asn/Ser is equivocal to increase cancer risk.

328 The novel HLA-A\*24 allele tagged the known intronic variant rs1264813 in *MICD*, and was  
329 similarly associated with increased risk of HPV(-) OPC (OR (95% CI)= 1.34 (1.18, 1.52),  
330  $p_{\text{meta}}=7.24 \times 10^{-6}$ ). Accuracy of the model including this allele was similar to the model including  
331 rs1264813, suggesting these signals convey similar risk (Figure 5c, Tables S12).

332 A novel haplotype was identified that tagged the known intronic variant, rs9268925 in *DRB9*,  
333 and was associated with decreased risk of OC (OR (95% CI)= 0.8 (0.73, 0.86),  $p_{\text{meta}}= 2.15 \times 10^{-$   
334 <sup>8</sup>). The haplotype, DRB1\*15:01-DQA1\*01:02-DQB1\*06:02, had a similar risk and similar  
335 model accuracy compared to the known variant, suggesting that this variant and the novel  
336 haplotype can be used interchangeable to measure this risk (Figure 5d). Two novel variants  
337 specific to the European ancestry were associated with risk of OC: DRB1 74Ala/Leu/Del (OR  
338 (95% CI)= 0.82 (0.77, 0.87),  $p=4.94 \times 10^{-10}$ ) and rs9267280 (OR (95% CI)= 1.32 (1.19, 1.47),  
339  $p=3.48 \times 10^{-7}$ ).

340 **Cross ancestry equivalent of established risk variants, including the well-known**  
341 **haplotype DRB1\*13:01-DQA1\*01:03-DQB1\*06:03**

342 The DRB1\*13:01-DQA1\*01:03-DQB1\*06:03 haplotype is well known to reduce risk of cervical  
343 cancer and HPV(+) OPC<sup>10,11,28</sup>. Notably, the two novel *DRB1* amino acid changes, DRB1  
344 37Asn/Ser and DRB1 233Thr, described here for risk of HPV(+) OPCs are within this  
345 haplotype (Figure 4a). To determine if the haplotype is completely represented by these amino  
346 acid changes, we replaced the amino acids with the full haplotype in the risk model for HPV(+)  
347 OPC (Figure 4c). Unexpectedly, the effect of HLA-B 67Cys/Ser/Tyr disappeared when  
348 including the haplotype, suggesting these are shared risk loci. When all three variants were  
349 replaced by the haplotype, the haplotype was independently associated with HPV(+) OPC risk  
350 (OR (95% CI)= 0.53 (0.43, 0.63),  $p_{\text{meta}}=1.76 \times 10^{-10}$ ), as described previously<sup>11</sup>. Importantly,  
351 model accuracy was highest for the model consisting of the original three amino acid changes  
352 compared to the haplotype, suggesting that the specific independent effects of the newly  
353 identified DRB1 37Asn/Ser, DRB1 233Thr, and possibly HLA-B 67Cys/Ser/Tyr underlie the  
354 effect of the DRB1\*13:01-DQA1\*01:03-DQB1\*06:03 haplotype. The importance of these  
355 amino acid changes is highlighted by their allele frequencies across populations, compared to  
356 the haplotype (Figure 4d). The allele frequency of the haplotype across genetic ancestries is  
357 low and ranges from 3% to 6%, while the frequency of the three amino acids across ancestries  
358 is much higher, ranging from 26% to 33%.

359 The novel rs2523679 variant, which decreases risk of HPV(+) OPC (OR (95% CI) = 0.63 (0.53,  
360 0.75),  $p_{\text{meta}}=2.26 \times 10^{-7}$ ), tags the established HLA-B\*15:01 ( $r^2=0.51$ ) and HLA-B 156Trp  
361 ( $r^2=0.51$ ) signals that were previously found in those of European ancestry. Here we show that  
362 while the effects of HLA-B\*15:01 and HLA-B 156Trp remain specific to European ancestry,  
363 rs2523679 confers a similar level of risk for both European and admixed populations, providing  
364 a cross-ancestral equivalent of this loci (Figure 4e). Other cross ancestral validated loci are  
365 described in Tables S10 and S11.

366 **Discussion**

367 Across the GWAS and HLA focused analyses, we identify 29 novel variants associated with  
368 risk of HNSCC. Due to increased power compared to previous GWA studies, we identified  
369 novel genetic variants including in *TP53* and *STING1* and validated known variants in *BRCA2*  
370 separately in LA and HPC, two under-studied cancer sites, as well as multiple novel signals in  
371 HPC, such as *GDF7*. Novel variants from fine mapping highlight key differences in HLA  
372 associations between HPV(+) OPCs, HPV(-) OPCs and OCs. Post-GWAS analyses, including

373 colocalisation and the use of harmonized individual level risk factor data enabled the  
374 investigation of variant function and variant-environment interactions.

375 A key finding was the identification of the low frequency rs78378222 variant located in the 3'  
376 UTR of *TP53*. This variant is an eQTL for *TP53* with the variant causing decreased expression  
377 in blood and a protective effect against overall HNSCC. This finding supports a previous  
378 candidate SNP study in a non-Hispanic white population assessing its effect on HNSCCs  
379 (OR=0.44, 95% CI: 0.24,0.79, p=0.008)<sup>29</sup>. Interestingly, while this variant is protective for  
380 HNSCCs and breast cancers<sup>30</sup>, it increases risk of skin basal cell carcinoma<sup>21</sup>, brain tumours,<sup>21</sup>  
381 colorectal adenocarcinoma<sup>21</sup>, esophageal SCC,<sup>31</sup> prostate cancers,<sup>21</sup> and neuroblastoma<sup>32</sup>.  
382 Mouse models developed by Deng and colleagues have demonstrated these contrasting  
383 effects<sup>30</sup>. Mice with the variant exhibited reduced tumor formation and increased survival rates  
384 for breast cancers but showed the opposite trend for gliomas. The authors demonstrated the  
385 rs78378222 variant compromises the microRNA (miR)-325 target site but creates a miR-382  
386 target site in the p53 mRNA. Both these microRNA molecules downregulate p53 but have  
387 differential tissue expressions. miR-325 expression is high in breast tissue meaning the variant  
388 p53 can escape the miR-325 downregulation, but brain tissue has high miR-382 expression,  
389 meaning the variant p53 has increased downregulation. Future work should investigate the  
390 effect of miR-325 regulation of p53 in head and neck tissues of carriers of rs78378222.

391 Two closely linked genetic variants were identified in 5q31, including the missense variant  
392 rs1131769 found in the cyclic dinucleotide (CDN) binding domain of *TMEM173* gene, of which  
393 the resultant STING1 protein detects viral DNA and bacterial CDNs to activate the host  
394 immune response in humans. Notably, this variant shows no association with HPV(+) OPC,  
395 but a consistent increased risk for all non-HPV cancer types. Both variants also showed  
396 evidence of eQTLs for *CTNNA1*, a gene in which germline genetic variants are known to cause  
397 Hereditary Diffuse Gastric Cancer<sup>33</sup>.

398 We were able to validate several known HNSCC risk variants and further investigate their  
399 interaction with major risk factors. rs11571833 has been linked with lung and upper  
400 aerodigestive tract cancers<sup>34</sup>; here we demonstrate this effect is largest in LA and HPC  
401 cancers. This variant, found in *BRCA2*, causes a 93 amino-acid deletion including the RAD51  
402 binding domain, important in the Fanconi Anaemia Pathway for double strand DNA repair, and  
403 is distinct to the highly penetrant familial *BRCA* mutations<sup>35</sup>. Previous literature suggests  
404 smoking is mainly implicated in the mechanism of action of rs11571833.<sup>27</sup> However, here we  
405 provide evidence cross ancestry and separately in the European and Mixed ancestry groups  
406 that this variant increases HPV-negative cancer risk with either the exposure of smoking or

407 drinking, and that there is no effect in never-smoking non-drinkers. This supports the theory  
408 that DNA repair to environmental factors is disrupted<sup>35</sup> and suggests the crucial DNA damage  
409 in HNSCC can be contributed to by alcohol use or smoking. In similar analyses, we show the  
410 well-known *ADH1B* variant rs1229984, confers a protective effect for OC which is strongest in  
411 non-drinking smokers, suggesting a mechanism through smoking as well as alcohol use. The  
412 *CHRNA5* variant, rs58365910, was identified as a suggestive association for risk of LA cancer.  
413 As expected, this variant only shows an effect in smokers, suggesting that it acts through its  
414 known effect on smoking heaviness.<sup>36</sup>

415 Through HLA fine mapping efforts, we identified 11 novel loci specific to the HLA region, of  
416 which eight were separately associated with risk of OC, HPV(+) OPC or HPV(-) OPC. Most of  
417 the class I loci were found in *HLA-B*, while most the class II loci were in *DRB1*. Given the  
418 dense, overlapping structure of the HLA region, we also identified functionally equivalent  
419 signals at the amino acid, allele, or haplotype level, enabling these data to support a variety  
420 of downstream applications requiring functional information.

421 A novel class II haplotype was identified for risk of OC, DRB1\*15:01-DQA1\*01:02-  
422 DQB1\*06:02. This haplotype has been found to reduce autoantibody development and  
423 abnormalities of metabolic traits, such as dysglycemia. As such, this haplotype was found to  
424 be protective against progression of type I diabetes (DM)<sup>37</sup>. The relevance of this finding is  
425 evidenced by a meta-analysis that found that individuals with DM have a higher risk of  
426 developing oral cancer<sup>38</sup>, potentially related to DM-related metabolic traits such as  
427 hypertension and dyslipidemia<sup>39</sup>. Nevertheless, a link between DM and OC remains  
428 inconsistent<sup>40-42</sup>. The OC-specific validated variant, rs4990036, is also associated with a non-  
429 HPV infection, varicella zoster<sup>43</sup>, highlighting that other infections may be important in cancer  
430 risk. This is especially important considering the oral microbiome as a potential emerging risk  
431 factor for oral cavity cancer.

432 The well-known haplotype, DRB1\*13:01-DQA1\*01:03-DQB1\*06:03, has been found to be  
433 protective against cervical cancer and HPV(+) OPC, highlighting its role in detecting HPV  
434 infection<sup>10,11,28</sup>. This haplotype is present at about 5% in the European ancestry and less  
435 common in other ancestries. We show here that the DRB1\*13:01-DQA1\*01:03-DQB1\*06:03  
436 haplotype is represented by the novel three novel amino acid changes identified in this work,  
437 DRB1 37Asn/Ser, DRB1 233Thr, and HLA-B 67Cys/Ser/Tyr. Notably however, the amino acid  
438 changes themselves more precisely estimate risk of HPV(+) OPC across ancestries and likely  
439 drive the effect of the haplotype across ancestries. The higher allele frequencies of the amino  
440 acids, ranging from 26% to 33%, allows for better detection of risk for HPV(+) OPC across

441 populations and might be easier to incorporate into screening strategies at the population  
442 level.

443 The novel intronic rs2523679 variant is a cross-ancestral equivalent of HLA-B\*15:01 and HLA-  
444 B 156Trp, two previously identified European-specific variants. This novel variant can now be  
445 used to evaluate risk of HPV(+) OPC across multiple ancestries, and highlights the importance  
446 of including non-European populations, even with limited sample size.

447 In this work, we were limited by the power from non-European populations, forcing us to  
448 combine multiple populations. Although this did provide additional power for discovery, it will  
449 have reduced the ability to identify variants specific to certain populations. Where variants  
450 were specific to non-European ancestries, we were able to assess these in the different  
451 populations but increased sample sizes from more diverse populations should still be seen as  
452 a priority in this field.

453 Although analysing all-site HNSCC can be beneficial, it must be remembered that these  
454 cancers are heterogeneous, and the subsite analyses provide a clearer picture of the genetic  
455 architecture of the conditions. Where we identify genetic variants in one site, we assess the  
456 effect of this variant across all subsites to assess the heterogeneity but despite the increased  
457 sample size in this study, there may still be limited power for discovery, especially in the less  
458 common subsites such as HPC.

459 In summary, in this HNSCC GWAS which includes diverse populations, we identify 29 novel  
460 genetic variants associated with HNSCC and its subsites, including rs78378222 in the *TP53*  
461 3' UTR which confers a 40% reduction in odds of developing overall HNSCC. We expand  
462 knowledge of the gene-environment relationship of *BRCA2* and *ADH1B* variants  
463 demonstrating their effects act through both smoking and alcohol use. Finally, through  
464 analyses focused on the HLA region, we highlight that although HPV(+) OPC, HPV(-) OPC  
465 and OC all show GWAS signal at 6p21, each subsite has distinct associations at the variant,  
466 amino acid and 4-digit allele level.

#### 467 **Data Availability Statement**

468 The full GWAS summary statistics generated from our meta-analyses will be made publicly  
469 accessible on (<https://gwas.mrcieu.ac.uk/>).

470 Additional datasets analyzed in this study are accessible through dbGaP  
471 (<https://www.ncbi.nlm.nih.gov/gap/>) as follows:

472 OncoArray Consortium - Lung Cancer Studies (dbGaP Study Accession: phs001273.v4.p2),

473 OncoArray: Oral and Pharynx Cancer (dbGaP Study Accession: phs001202.v2.p1),  
474 National Cancer Institute (NCI) Head and Neck Cancer Study conducted on the  
475 HumanOmniExpress-12v1.0 array (dbGaP Study Accession: phs001173.v1.p1),  
476 Genome-Wide Association Study of Oral Cavity, Pharynx, and Larynx Cancers in European,  
477 North, and South American populations (dbGaP Study Accession: phs002503.v1.p1).  
478 Data from the UK Biobank and ALSPAC consortium, are available through their respective  
479 access protocols.

#### 480 **Code availability statement**

481 This study did not employ any custom code. Instead, it utilized publicly available software tools  
482 for genetic analyses, which are cited throughout the manuscript and reporting summary.

#### 483 **References**

- 484 1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of  
485 Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer*  
486 *J Clin* **71**, 209–249 (2021).
- 487 2. Johnson, D. E. *et al.* Head and neck squamous cell carcinoma. *Nature Reviews*  
488 *Disease Primers* 2020 6:1 **6**, 1–22 (2020).
- 489 3. Lubin, J. H. *et al.* An examination of male and female odds ratios by BMI,  
490 cigarette smoking, and alcohol consumption for cancers of the oral cavity,  
491 pharynx, and larynx in pooled data from 15 case-control studies. *Cancer Causes*  
492 *and Control* **22**, 1217–1231 (2011).
- 493 4. Thomas, S. J., Penfold, C. M., Waylen, A. & Ness, A. R. The changing aetiology  
494 of head and neck squamous cell cancer: A tale of three cancers? *Clinical*  
495 *Otolaryngology* vol. 43 999–1003 (2018).
- 496 5. Hobbs, C. G. L. *et al.* Human papillomavirus and head and neck cancer: a  
497 systematic review and meta-analysis. *Clinical Otolaryngology* **31**, 259–266  
498 (2006).
- 499 6. Gillison, M. L., Chaturvedi, A. K., Anderson, W. F. & Fakhry, C. Epidemiology of  
500 Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma. *J*  
501 *Clin Oncol* **33**, 3235–3242 (2015).
- 502 7. Chaturvedi, A. K. *et al.* Human papillomavirus and rising oropharyngeal cancer  
503 incidence in the United States. *J Clin Oncol* **29**, 4294–4301 (2011).

- 504 8. Jamieson, L. M. *et al.* Cohort profile: indigenous human papillomavirus and  
505 oropharyngeal squamous cell carcinoma study - a prospective longitudinal  
506 cohort. *BMJ Open* **11**, e046928 (2021).
- 507 9. WHO Classification of Tumours Editorial Board. *Head and Neck Tumours: WHO*  
508 *Classification of Tumours*. vol. 9 (International Agency for Research on Cancer;  
509 Forthcoming., 2024).
- 510 10. Ferreiro-Iglesias, A. *et al.* Germline determinants of humoral immune response  
511 to HPV-16 protect against oropharyngeal cancer. *Nat Commun* **12**, (2021).
- 512 11. Lesseur, C. *et al.* Genome-wide association analyses identify new susceptibility  
513 loci for oral cavity and pharyngeal cancer. *Nat Genet* **48**, 1544–1550 (2016).
- 514 12. McKay, J. D. *et al.* A Genome-Wide Association Study of Upper Aerodigestive  
515 Tract Cancers Conducted within the INHANCE Consortium. *PLoS Genet* **7**,  
516 e1001333 (2011).
- 517 13. Lesseur, C. *et al.* Genome-wide association meta-analysis identifies pleiotropic  
518 risk loci for aerodigestive squamous cell cancers. *PLoS Genet* **17**, e1009254  
519 (2021).
- 520 14. Shete, S. *et al.* A Genome-Wide Association Study Identifies Two Novel  
521 Susceptible Regions for Squamous Cell Carcinoma of the Head and Neck.  
522 *Cancer Res* (2020) doi:10.1158/0008-5472.CAN-19-2360.
- 523 15. Elmusrati, A., Wang, J. & Wang, C. Y. Tumor microenvironment and immune  
524 evasion in head and neck squamous cell carcinoma. *Int J Oral Sci* **13**, (2021).
- 525 16. Zhang, Z. *et al.* Polymorphisms in the PVT1 Gene and Susceptibility to the Lung  
526 Cancer in a Chinese Northeast Population: a Case-control Study. *J Cancer* **11**,  
527 468–478 (2020).
- 528 17. Song, N. *et al.* Evaluation of gene-environment interactions for colorectal cancer  
529 susceptibility loci using case-only and case-control designs. *BMC Cancer* **19**, 1–  
530 10 (2019).
- 531 18. Lesseur, C. *et al.* A case-control study of polymorphisms in xenobiotic and  
532 arsenic metabolism genes and arsenic-related bladder cancer in New  
533 Hampshire. *Toxicol Lett* **210**, 100–106 (2012).
- 534 19. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER  
535 negative-specific breast cancer risk loci. *Nat Genet* **45**, (2013).



- 536 20. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using  
537 the iCOGS custom genotyping array. *Nat Genet* **45**, (2013).
- 538 21. Stacey, S. N. *et al.* A germline variant in the TP53 polyadenylation signal confers  
539 cancer susceptibility. *Nat Genet* **43**, 1098–1103 (2011).
- 540 22. Zaidi, A. A., Verma, A., Morse, C., Ritchie, M. D. & Mathieson, I. The genetic and  
541 phenotypic correlates of mtDNA copy number in a multi-ancestry cohort. *HGG*  
542 *Adv* **4**, (2023).
- 543 23. Hägg, S., Jylhävä, J., Wang, Y., Czene, K. & Grassmann, F. Deciphering the  
544 genetic and epidemiological landscape of mitochondrial DNA abundance. *Hum*  
545 *Genet* **140**, 849–861 (2021).
- 546 24. Guo, H., Cao, W., Zhu, Y., Li, T. & Hu, B. A genome-wide cross-cancer meta-  
547 analysis highlights the shared genetic links of five solid cancers. *Front Microbiol*  
548 **14**, (2023).
- 549 25. Drobni, P., Näslund, J. & Evander, M. Lactoferrin inhibits human papillomavirus  
550 binding and uptake in vitro. *Antiviral Res* **64**, 63–68 (2004).
- 551 26. Bukowska-Oško, I. *et al.* Lactoferrin as a Human Genome “Guardian”—An  
552 Overall Point of View. *International Journal of Molecular Sciences 2022, Vol. 23,*  
553 *Page 5248* **23**, 5248 (2022).
- 554 27. Delahaye-Sourdeix, M. *et al.* A rare truncating BRCA2 variant and genetic  
555 susceptibility to upper aerodigestive tract cancer. *J Natl Cancer Inst* **107**, (2015).
- 556 28. Chen, D. *et al.* Genome-wide association study of susceptibility loci for cervical  
557 cancer. *J Natl Cancer Inst* **105**, 624–633 (2013).
- 558 29. Guan, X., Wang, L. E., Liu, Z., Sturgis, E. M. & Wei, Q. Association between a  
559 rare novel TP53 variant (rs78378222) and melanoma, squamous cell carcinoma  
560 of head and neck and lung cancer susceptibility in non-Hispanic Whites. *J Cell*  
561 *Mol Med* **17**, 873–878 (2013).
- 562 30. Deng, Q. *et al.* Tissue-specific microRNA expression alters cancer susceptibility  
563 conferred by a TP53 noncoding variant. *Nature Communications 2019 10:1* **10**,  
564 1–13 (2019).
- 565 31. Zhou, L., Yuan, Q. & Yang, M. A functional germline variant in the P53  
566 polyadenylation signal and risk of esophageal squamous cell carcinoma. *Gene*  
567 **506**, 295–297 (2012).

- 568 32. Diskin, S. J. *et al.* Rare variants in TP53 and susceptibility to neuroblastoma. *J*  
569 *Natl Cancer Inst* **106**, (2014).
- 570 33. Lobo, S. *et al.* Cancer predisposition and germline CTNNA1 variants. *Eur J Med*  
571 *Genet* **64**, (2021).
- 572 34. Delahaye-Sourdeix, M. *et al.* A rare truncating BRCA2 variant and genetic  
573 susceptibility to upper aerodigestive tract cancer. *J Natl Cancer Inst* **107**, (2015).
- 574 35. Rafnar, T. *et al.* Association of brca2 k3326\* with small cell lung cancer and  
575 squamous cell cancer of the skin. *J Natl Cancer Inst* **110**, (2018).
- 576 36. Ware, J. J., Van den bree, M. B. M. & Munafò, M. R. Association of the CHRNA5-  
577 A3-B4 Gene Cluster With Heaviness of Smoking: A Meta-Analysis. *Nicotine &*  
578 *Tobacco Research* **13**, 1167 (2011).
- 579 37. Pugliese, A. *et al.* HLA-DRB1\*15:01-DQA1\*01:02-DQB1\*06:02 Haplotype  
580 Protects Autoantibody-Positive Relatives From Type 1 Diabetes Throughout the  
581 Stages of Disease Progression. *Diabetes* **65**, 1109–1119 (2016).
- 582 38. Ramos-Garcia, P., Roca-Rodriguez, M. del M., Aguilar-Diosdado, M. &  
583 Gonzalez-Moles, M. A. Diabetes mellitus and oral cancer/oral potentially  
584 malignant disorders: A systematic review and meta-analysis. *Oral Dis* **27**, 404–  
585 421 (2021).
- 586 39. Tseng, K. S., Lin, C., Lin, Y. S. & Weng, S. F. Risk of head and neck cancer in  
587 patients with diabetes mellitus: a retrospective cohort study in Taiwan. *JAMA*  
588 *Otolaryngol Head Neck Surg* **140**, 746–753 (2014).
- 589 40. Zhou, X. H. *et al.* Diabetes, prediabetes and cancer mortality. *Diabetologia* **53**,  
590 1867–1876 (2010).
- 591 41. Lo, S. F. *et al.* Modest increase in risk of specific types of cancer types in type 2  
592 diabetes mellitus patients. *Int J Cancer* **132**, 182–188 (2013).
- 593 42. Stott-Miller, M. *et al.* History of diabetes and risk of head and neck cancer: a  
594 pooled analysis from the international head and neck cancer epidemiology  
595 consortium. *Cancer Epidemiol Biomarkers Prev* **21**, 294–304 (2012).
- 596 43. Kachuri, L. *et al.* The landscape of host genetic factors involved in immune  
597 response to common viral infections. *Genome Medicine* **2020 12:1** **12**, 1–18  
598 (2020).
- 599 44. Hibbert, J., Halec, G., Baaken, D., Waterboer, T. & Brenner, N. Sensitivity and  
600 specificity of human papillomavirus (Hpv) 16 early antigen serology for hpv-

- 601 driven oropharyngeal cancer: A systematic literature review and meta-analysis.  
602 *Cancers* vol. 13 Preprint at <https://doi.org/10.3390/cancers13123010> (2021).
- 603 45. Meyer HV. plinkQC: Genotype quality control in genetic association studies.  
604 Preprint at <https://doi.org/10.5281/zenodo.3934294> (2020).
- 605 46. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger  
606 and richer datasets. *Gigascience* **4**, 7 (2015).
- 607 47. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed  
608 Program. *Nature* 2021 590:7845 **590**, 290–299 (2021).
- 609 48. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat*  
610 *Genet* **48**, 1284–1287 (2016).
- 611 49. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and  
612 children: ALSPAC mothers cohort. *Int J Epidemiol* **42**, 97–110 (2013).
- 613 50. Jones, R. W. *et al.* A new human genetic resource: a DNA bank established as  
614 part of the Avon longitudinal study of pregnancy and childhood (ALSPAC). *Eur J*  
615 *Hum Genet* **8**, 653–660 (2000).
- 616 51. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association  
617 studies. *Bioinformatics* **26**, 2867–2873 (2010).
- 618 52. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of  
619 ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
- 620 53. 1000 Genomes Project Consortium *et al.* A global reference for human genetic  
621 variation. *Nature* (2015) doi:10.1038/nature15393.
- 622 54. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of  
623 genomewide association scans. *BIOINFORMATICS APPLICATIONS NOTE* **26**,  
624 2190–2191 (2010).
- 625 55. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population  
626 diversity enables multi-ancestry fine-mapping in HIV host response.  
627 doi:10.1038/s41588-021-00935-7.
- 628 56. Okada, Y. *et al.* Fine mapping major histocompatibility complex associations in  
629 psoriasis and its clinical subtypes. *Am J Hum Genet* **95**, 162–172 (2014).
- 630 57. Yang, J. A. *et al.* GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J*  
631 *Hum Genet* **88**, 76–82 (2011).

- 632            58.    Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing  
633                            heritability for disease from genome-wide association studies. *Am J Hum Genet*  
634                            **88**, 294–305 (2011).
- 635            59.    Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of  
636                            Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, (2014).
- 637            60.    Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of  
638                            genetic loci and polygenic scores that regulate blood gene expression. *Nat*  
639                            *Genet* **53**, 1300–1310 (2021).
- 640            61.    Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across  
641                            human tissues. *Science (1979)* **369**, (2020).
- 642            62.    Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights  
643                            into the genetic etiology of tobacco and alcohol use. *Nature Genetics* vol. 51  
644                            Preprint at <https://doi.org/10.1038/s41588-018-0307-5> (2019).
- 645
- 646

## 647 **Online Methods**

### 648 **Study Design and Populations**

649 Individual level data came from 18 studies across 23 countries in Europe, Middle East, North  
650 America, South America, and South Asia, and 9 genotyping arrays (Table S1). Informed  
651 consent and ethical approval for genotyping was obtained under each individual study. An  
652 overall ethical approval for this analysis was obtained from the IARC Ethics Committee (IEC,  
653 19-38). Data on demographics (sex, age, country), diagnosis (TNM status, year of diagnosis,  
654 ICD code -7th edition), HPV status (HPV16E6 serology, P16 immunohistochemistry (IHC), and  
655 HPV DNA in situ hybridisation (ISH)) and self-reported behaviors (smoking status, packyears,  
656 and drinking status) were collated and harmonized across all study participants. Eligible sites  
657 for inclusion consisted of the oral cavity (C00.3, C00.4, C00.5, C00.6, C00.8, C00.9, C02.0–  
658 C02.9 (except C02.4 and C02.8), C03.0–C03.9, C04.0–C04.9, C05.0–C06 (except C05.1,  
659 C05.2)); oropharynx (C01-C01.9, C02.4, C05.1, C05.2, and C09.0–C10.9); hypopharynx  
660 (C12.0-C13.0); larynx (C32); and unknown primary site/overlapping/not otherwise specified  
661 (NOS) sites (C14, C05.8, C02.8, C76.0). Base of tongue (C01) and tonsils (C09) were  
662 grouped with oropharynx as these sites are frequently driven by HPV16. For studies with  
663 available information on HPV infection for OPC tumors, the HPV status provided by the centre  
664 was used (P16 status, HPV DNA ISH, or HPV serology). When information from various  
665 methods was available, a positive HPV status was determined by the presence of the HPV16  
666 E6 antibody in serology. If serology data were absent, dual positivity of p16 and HPV DNA ISH  
667 was classified as HPV positive (HPV(+)), while dual negativity of p16 and HPV DNA ISH was  
668 classified as HPV negative (HPV(-)). Any other combinations of test results were considered  
669 as “not available”<sup>44</sup>.

670 Nineteen studies were included here with either multi-center case-control, cohort, or clinical  
671 trial study designs. Previously generated data was either downloaded from dbGap, requested  
672 through controlled access from relevant consortia, or contributed by the study PI and  
673 contributed 10,404 cases and 34,596 controls. New genotyping data was generated for 8,669  
674 cases and 4261 controls and were not included in any previous GWAS. All study details,  
675 including data sources, dbGap accession numbers and case control distributions across  
676 subsites can be found in Table S1.

### 677 **Genotype quality control and imputation**

678 A flow diagram detailing the preparation of the genetic data can be found in the supplementary  
679 material (Figure S10). Genotypes were generated from nine different genotyping arrays (Table

680 S1). All newly generated genotype data was called using GenomeStudio (Illumina, 2014).  
681 Quality control steps were conducted within each array. All genotype data were converted to  
682 genome build 38, using the LiftOver program (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>) to  
683 convert from previous builds. Genotype data was checked and corrected for consistency of  
684 strand, positions and reference alleles. Quality control was conducted using the PlinkQC  
685 package<sup>45</sup> in R, utilising PLINK 1.9<sup>46</sup>. Samples were filtered for sex mismatch (males with  
686 SNP sex <0.8; females with SNP sex >0.2), missingness (>3%), heterozygosity (>3 standard  
687 deviations from mean) and cryptic relatedness (identity-by-descent > 0.185). Variants were  
688 filtered for genotype missingness (>1%), deviation from Hardy Weinberg equilibrium ( $p < 1 \times 10^{-5}$ )  
689 and minor allele count (<20). The number of samples and variants removed at each QC  
690 step is provided in Table S2 and Table S13. All arrays were imputed to the TOPMED imputation  
691 panel<sup>47</sup> separately using the TOPMED Imputation server<sup>48</sup>.

692 To increase the number of controls comparable to the participants in the HN5000 study, 17,815  
693 additional participants (including known related individuals) were included from the Avon  
694 Longitudinal Study of Parents and Children (ALSPAC), which had been previously genotyped  
695 (Table S1)<sup>49,50</sup>. To account for potential batch effects between the HN5000 study (Infinium  
696 Global Screening Array [GSA]) and additional ALSPAC controls (Illumina 550 Quad, Illumina  
697 660W Quad), a double imputation approach was applied (Supplementary Note 1). Briefly, GSA  
698 HN5000 cases and the additional controls were imputed to the TOPMED reference panel  
699 separately as detailed above. Following this step, variants which were (i) genotyped on both  
700 arrays, (ii) genotyped on the GSA with high quality imputation ( $R^2$  score >0.9) on the ALSPAC  
701 array and, (iii) genotyped on the ALSPAC array with high quality imputation ( $R^2$  score >0.9) on  
702 the GSA were selected. These variants were merged across the two arrays, converted to 'best-  
703 guess' genotypes and then included in a second joint imputation to the TOPMED reference  
704 panel. This method allowed high quality imputation of both datasets. To address concerns  
705 about batch effects between cases and controls genotyped separately, 405 ALSPAC controls  
706 were also genotyped on the GSA alongside the HN5000 cases. This enabled sensitivity  
707 analyses to account for potential batch effects.

## 708 **Genetic Ancestry stratification**

709 Following the imputation process, markers from each imputation batch were filtered based on  
710 an imputation score of  $R^2 > 0.8$  and merged across imputation batch and chromosome.  
711 Markers were filtered for a call rate  $\geq 0.98$  and minor allele frequency (MAF)  $\geq 1\%$ . The major  
712 histocompatibility complex (MHC) region was removed, and the remaining markers were  
713 pruned for independent variants using linkage disequilibrium (LD) with a squared correlation

714 ( $r^2$ ) threshold of  $< 0.2$ . This set of markers ( $N=697,099$ ) was utilized to compute kinship  
715 estimates between Individuals using the KING-robust kinship estimator<sup>51</sup> in PLINK 2.0<sup>46</sup>. The  
716 KING-robust method is specifically designed to be robust to population structure and  
717 admixture. It calculates kinship coefficients without being biased by the fact that certain  
718 populations may have different allele frequencies. In addition to the removal of 6,679 known  
719 related individuals from the ALSPAC study, a kinship cutoff of  $> 0.0884$  was applied to exclude  
720 unexpected duplicates and individuals related at the second degree or closer. This cutoff is  
721 based on the geometric mean of the theoretical values for second- and third-degree kinship,  
722 as outlined in the manual. Selection of related individuals or duplicates were prioritized based  
723 on either disease status (favouring cases over control) or array type (favouring newer arrays  
724 over older ones). After this process, 3,441 individuals were excluded from the analysis. The  
725 remaining 58,625 individuals were classified into genetic ancestries using supervised  
726 ADMIXTURE analysis (ADMIXTURE 1.3<sup>52</sup>) with 75,164 common markers retained after  
727 quality control steps (Figure S11). This assigns a percentage probability for belonging to each  
728 of the reference super-populations in the 1000 Genomes Project ( $N=2,504$ )<sup>53</sup>. We assigned  
729 individuals to a dominant genetic ancestry if their probability was  $\geq 70\%$  to any reference super-  
730 population. Of all individuals, 48,029 (83%) had a dominant genetic ancestry while the  
731 remainder were classified as admixed. The distribution of individuals with a dominant genetic  
732 ancestry was as follows: 80.2% European (EUR), 0.1% Admixed Americans (AMR), 1.2%  
733 Africans (AFR), 1.3% South Asians (SAS), and 0.2% East Asians (EAS). The remaining 17%  
734 were not able to be classified with a dominant genetic ancestry and were grouped as  
735 "admixed". To improve statistical power to detect risk loci across the relatively small sample  
736 sizes of non-European genetic ancestries, all five (AMR, AFR, SAS, EAS and admixed) were  
737 merged to create a "Mixed" group ( $n = 11,462$ ) (Figure S12a, S12b). Genome-Wide  
738 Association Studies (GWAS) were conducted separately in the European and Mixed ancestry  
739 samples and meta-analysed (see later). Principal Component Analysis (PCA) was carried out  
740 within each ancestral sample (European and Mixed) to assess population substructure and  
741 for covariate adjustment in GWAS (Figure S13). For HLA fine-mapping analyses, a slightly  
742 different approach was required due to the region's high LD and highly correlated variants.  
743 Additionally, the HLA region is more susceptible to population substructure, making it  
744 challenging to identify causal variants that are consistent across ancestries. Therefore, for fine  
745 mapping, samples were grouped according to their dominant genetic ancestry ( $> 70\%$ ) (EUR,  
746 AFR, and SAS) or admixed. Based on the homogeneous clustering identified through PCA  
747 (Figure S14), the samples from Iran were separated to Middle Eastern (ME) ancestry. Small  
748 size numbers (Case/Control  $< 50$ ) of genetic ancestries (AMR and EAS) were merged into  
749 admixed. For each sample, PCA identified informative principal components (PCs) that  
750 showed significant associations ( $p < 0.05$ ) with case-control status after adjusting for sex and

751 imputation batch. These informative PCs, along with sex and imputation batch, were included  
752 as model covariates in the GWAS analysis.

### 753 **Association, Meta- and Conditional Analysis**

754 Across the 9 arrays and 19 studies, there were several considerations in how to adjust for  
755 batch effects. Some studies, such as ARCAGE, were split across different arrays, such as the  
756 Oncoarray and AllofUs array. For other studies, such as UKBiobank, several arrays were used  
757 (UKBiLEVE and AffymetricUKB) (Table S1). Finally, HN5000 and ALSPAC differed in their  
758 imputation as the 'double imputation' method was used. Therefore, a 'Batch' variable was  
759 created to represent the combination of studies, arrays and imputation approaches that could  
760 contribute to batch effects. To evaluate the potential impact of these different batches in the  
761 regression models, we conducted a sensitivity analysis by running GWAS within each batch  
762 and assessed heterogeneity using METAL<sup>54</sup>. We excluded markers with a heterogeneity p-  
763 value  $<5 \times 10^{-8}$ , resulting in the removal of 137 markers in the European sample GWAS.

764 Association analysis was conducted separately for all sites combined and for each HNSCC  
765 subsite in the European and Mixed samples using PLINK. The results were then meta-  
766 analysed with METAL<sup>54</sup> using a fixed effects model to identify cross-ancestral loci. There was  
767 minimal inflation after adjustment for informative PCs in most analyses ( $\lambda$ ) ranging  
768 from 1.00 to 1.03). However, the HPV(+) and HPV(-) OPC analyses for the Mixed group did  
769 show evidence of inflation (HPV(+) OPC:  $\lambda=1.12$ ; HPV(-) OPC:  $\lambda=1.20$ ) (Figure S15).  
770 Consequently, rather than a meta-analysis, the GWAS analysis for OPC was conducted only  
771 in the European sample with consistency of top SNPs assessed separately in the Mixed  
772 sample. For all other subsites, loci that achieved  $p < 5 \times 10^{-8}$  in the meta-analysis were referred  
773 to as cross-ancestral. This threshold was selected as it is equivalent to a standard Bonferonni  
774 correction for one million independent tests. Loci satisfying  $p < 5 \times 10^{-8}$  within each ancestral  
775 sample which 1) were not significant in the meta-analysis and 2) showed no attenuation upon  
776 conditional analysis of nearby lead cross-ancestral SNPs and therefore considered to be  
777 independent from the cross-ancestral SNP, were hereby referred to as ancestral-specific  
778 (Table S14). Where these existed in the Mixed ancestry sample, further stratification into the  
779 five dominant genetic ancestries was performed. Regional association plots were generated  
780 using Locus Zoom (<https://my.locuszoom.org/>).

### 781 **HLA fine mapping**

782 Variants that were directly genotyped in chromosome 6 were extracted from genotyping data  
783 of all arrays and standardized to hg19 using LiftOver. Due to restrictions in data access from



784 ALSPAC, additional data from UK Biobank was used to replace ALSPAC for double imputation  
785 with HN5000 as described above. Per variant QC was conducted by deduplication of SNP  
786 data, strand alignment, removal of palindromic variants (i.e., SNPs with A/T or G/C alleles),  
787 removal of poor-quality variants with missingness threshold of 10% and Hardy-Weinberg  
788 equilibrium threshold of  $1 \times 10^{-10}$ . Sample QC was conducted after the removal of samples with  
789 high missingness rates, outlier heterozygosity, discordant sex information, and genetically  
790 identical samples. A flow diagram of QC steps for the HLA fine mapping is provided in Figure  
791 S16.

792 The HLA region (Chromosome 6:28Mb-34Mb) was imputed for SNPs and classical HLA class  
793 I and II alleles using the Michigan imputation server with the most recent HLA Multi-ethnic  
794 reference panel (Four-digit Multi-ethnic HLA v2)<sup>55</sup>. Only high-quality SNPs, alleles or amino-  
795 acid residues were included in the analysis (imputation  $r^2 > 0.95$ ). The final set of imputed  
796 variants used in association analyses were of high quality; 91% of the variants and 71% of the  
797 less common variants (MAF < 0.05) had imputation  $R^2 \geq 0.95$ . HLA-wide association analysis  
798 was conducted controlling for sex, informative PCs, and imputation batch (described above),  
799 and meta-analyzed with a random effect model using PLINK<sup>46</sup> to identify cross-ancestral  
800 variants. Any genetic ancestries with fewer than 50 samples were excluded from meta-  
801 analyses due to power. Stepwise conditional analysis was conducted to identify independent  
802 variants within each ancestry where variants with the lowest p-value after each round were  
803 added to the subsequent model and the analysis was repeated until no further variants met  
804 the significance threshold. As HLA fine mapping was conducted independently from GWAS, a  
805 probability threshold was set to  $2.4 \times 10^{-6}$ . This was based on the total number of imputed HLA  
806 variants (0.05/20,762), which included SNPs, amino acid variants, and classical HLA alleles  
807 after quality control as described previously<sup>56</sup>.

808 To identify haplotypes associated with risk within each subsite that were linked to the top novel  
809 variants identified from fine mapping, the haplo.stats package v.1.9.5.1 in R was applied to  
810 identify combinations of HLA 4-digit alleles within each population. The haplo.em and  
811 haplo.glm algorithms identified haplotype candidates in each population with a minimum  
812 haplotype frequency threshold set at 0.01 in comparison to the most common haplotype within  
813 the ancestry. Haplotype candidates that were in high LD ( $r^2 > 0.8$ ) with variants from fine  
814 mapping were then tested for association with risk using the meta-analysis approach to  
815 determine if they conferred similar risk compared to their variant counterparts.

816 **Testing for Independence and functional equivalents of lead variants**

817 Variants identified in each HNSCC subsite analysis from the GWAS, and fine mapping were  
818 compared across subsites to evaluate whether they were linked or independent. This was also  
819 performed to define variants that were novel compared to previously reported signals and to  
820 determine overlapping signals between cross-ancestral and population-specific variants. LD  
821 was measured by  $r^2$  using PLINK 1.9<sup>46</sup> within the overall dataset. If  $LD > 0.3$ , then conditional  
822 analysis was performed to evaluate if the significance of the variant of interest attenuated to  
823 lower than  $2.4 \times 10^{-6}$ . If both criteria were met, variants were considered to be dependent.

824 To determine functional equivalents of the variants identified through fine mapping, amino acid  
825 changes, alleles and haplotypes that were in moderate to high LD ( $r^2 > 0.5$ ) with lead novel  
826 variants were further evaluated. Effect sizes and significance levels were compared when  
827 replacing the lead variant with the related variant in the fully adjusted cross-ancestral model.  
828 Bayesian Information Criterion (BIC) were then evaluated to compare the model fit of the  
829 original model with the lead variants identified from fine mapping to the model with the related  
830 variant replacing the original lead variant. Every permutation of variants was considered to  
831 determine if one variant could replace by another and still provide the same information as the  
832 original lead variant.

### 833 **Stratified analyses**

834 For each independent top hit identified in GWAS and HLA fine mapping, the analysis was  
835 repeated, stratified by sex, smoking status, drinking status, geographic region, and within all  
836 cancer subsites separately. The effects across strata were assessed for heterogeneity using  
837  $\chi^2$ -based Q test (Cochran's Q test) using R (v4.1.2). Further stratification for specific variants  
838 related to smoking and alcohol was conducted in non-HPV related cancers. This assessed  
839 effects in never-smoking non-drinkers, smoking non-drinkers, never-smoking drinkers and  
840 ever-smoking drinkers to assess the independence of these risk factors where data was  
841 available. Results were presented in forest plots (Figures S7, S9).

### 842 **Heritability and genetic correlation**

843 SNP based heritability was estimated in the European and Mixed ancestry samples using the  
844 Genome-based restricted maximum likelihood (GREML) method in GCTA<sup>57</sup>. Imputed genetic  
845 data was used, variants with  $MAF < 0.01$  and Hardy Weinberg equilibrium  $p < 0.05$  were  
846 removed, as suggested for case-control data in Lee *et al.* 2011<sup>58</sup>. Univariate GREML was used  
847 to estimate heritability and was transformed onto the liability scale using global prevalence  
848 estimates from GLOBOCAN<sup>1</sup>. Heritability estimates in the Mixed ancestry sample are not

849 presented in the main manuscript due to the heterogeneous nature of these samples which  
850 make estimates of heritability unreliable. These are provided in Table S5 for completeness.

### 851 **Colocalisation of GWAS and eQTL mapping**

852 Colocalisation of genetic associations between all identified top hit variants (from GWAS  
853 analyses and fine mapping) and their gene expression and related traits was calculated using  
854 default LDs and a window size of  $\pm 75$  kb using the COLOC package<sup>59</sup>. All colocalisation  
855 analyses were conducted using HNSCC data of European ancestry. Expression quantitative  
856 trait loci (eQTLs) in whole blood were obtained from the eQTLGen Consortium<sup>60</sup> due to its role  
857 in immune response and systemic inflammation. eQTLs in esophagus and lung tissue were  
858 sourced from the Genotype-Tissue Expression (GTEx) project (v8)<sup>61</sup>, given their anatomical  
859 proximity and shared risk factors, such as tobacco and alcohol exposure. In the analyses, we  
860 considered eQTLs coinciding with genomic loci identified in this study at  $p < 3.9 \times 10^{-10}$  in whole  
861 blood data;  $p < 2.5 \times 10^{-7}$  in tissue data to be considered as significant). Summary statistics from  
862 GWAS for smoking and alcohol consumption behaviors were sourced from the GWAS &  
863 Sequencing Consortium on Alcohol and Nicotine Use (GSCAN)<sup>62</sup>. The analysis considers the  
864 posterior probability of colocalisation for a single shared variant responsible for the  
865 associations in both traits (posterior probability for hypothesis 4 (PP4)), values over 0.7 were  
866 considered strong evidence of colocalisation. Where the lead variant was not available in the  
867 LD reference panel required for COLOC, the variant with the highest LD was used instead.

### 868 **Technical validation**

869 For the technical validation of the imputed *TP53* variant, we utilized a Taqman assay to  
870 genotype this specific variant in a subset of samples from the Central and Eastern European  
871 Study (CEE) and ARCAGE studies. Individuals removed from the GWAS in QC steps or those  
872 with technical issues during the Taqman assays, e.g. failure to amplify, were removed resulting  
873 in 2,370 samples where consistency could be assessed. Overall concordance and non-  
874 reference discordance were calculated.

## 875 HEADSpAcE Consortium Member Acknowledgements

876 Adam R<sup>1</sup>, Agudo A<sup>2</sup>, Alibhai S<sup>3</sup>, AlWaheidi SF<sup>4</sup>, Angel Pavon M<sup>2</sup>, Anwar N<sup>5</sup>, Arantes  
877 P<sup>6</sup>, Arguello L<sup>7</sup>, Avello Y<sup>8</sup>, Avondet L<sup>1</sup>, Baldión-Elorza AM<sup>8</sup>, Batista Daniel C<sup>9</sup>, Beraldi  
878 B<sup>10</sup>, Berenstein B<sup>11</sup>, Bernal P<sup>12</sup>, Bernardino Rodrigues N<sup>13</sup>, Bilic Zimmermann J<sup>2</sup>, Botta  
879 MG<sup>6</sup>, Bouvard L<sup>4</sup>, Brenes J<sup>2</sup>, Brenner N<sup>14</sup>, Brentisci C<sup>15</sup>, Burtica C<sup>8</sup>, Cabañas ML<sup>16</sup>,  
880 Cantor E<sup>17</sup>, Carvalho RS<sup>18</sup>, Carvalho A L<sup>19</sup>, Chiusa L<sup>20</sup>, Chopard P<sup>4</sup>, Chundriger Q<sup>21</sup>,  
881 Clavero O<sup>2</sup>, Costa I<sup>22</sup>, Creaney G<sup>23</sup>, Cuffini C<sup>24</sup>, Dias TC<sup>18</sup>, Duccini de Souza E<sup>10</sup>,  
882 durant I C<sup>25</sup>, Escallón A<sup>26</sup>, Fernandes G<sup>6</sup>, Fervers B<sup>27</sup>, Fiano V<sup>28</sup>, Firme Figueira F<sup>13</sup>,  
883 Furbino Villefort R<sup>13</sup>, Gangemi M<sup>15</sup>, Garzino-Demo P<sup>29</sup>, Gholipour M<sup>30</sup>, Giglio R<sup>11</sup>,  
884 Goulart MA<sup>31</sup>, Graça Sant'Anna J<sup>9</sup>, Grega M<sup>32</sup>, Gregório Có A<sup>9</sup>, Guasch A<sup>2</sup>, Hakim  
885 JA<sup>26</sup>, Hayes DN<sup>33</sup>, Homero de Sá Santos M<sup>10</sup>, Hurley K<sup>34</sup>, Insfran M<sup>35</sup>, Iorio GC<sup>36</sup>,  
886 Iqbaluddin Siddiqui M<sup>37</sup>, Johannsen J<sup>38</sup>, Kaña M<sup>39</sup>, Klussmann J<sup>38</sup>, Legal E<sup>40</sup>, Lenzi  
887 J<sup>10</sup>, Luiz Dias F<sup>22</sup>, Lyra González I<sup>41</sup>, Machado Zorzaneli W<sup>9</sup>, Mai Rocha R<sup>10</sup>, Maños  
888 M<sup>2</sup>, Marinho de Abreu P<sup>9</sup>, Marzban M<sup>42,43</sup>, McCaul J<sup>44</sup>, McMahon AD<sup>31</sup>, Mena C<sup>40</sup>,  
889 Mendonça EF<sup>45</sup>, Mendoza L<sup>35</sup>, Meza L<sup>35</sup>, Michels B<sup>14</sup>, Mineiro MS<sup>46</sup>, Moccia C<sup>28</sup>,  
890 Mongelos P<sup>35</sup>, Montealegre-Páez AL<sup>47</sup>, Morey Cortes F<sup>2</sup>, Muñoz A<sup>48</sup>, Ness A<sup>34</sup>, Neves  
891 AB<sup>6</sup>, Oliva M<sup>2</sup>, Oliveira J<sup>49</sup>, Ortiz H<sup>7</sup>, Ortiz J<sup>40</sup>, Osorio M<sup>40</sup>, Ospina V<sup>17</sup>, Ostellino O<sup>50</sup>,  
892 Palau M<sup>8</sup>, Paterson C<sup>51</sup>, Paytubi Casabona S<sup>2</sup>, Pecorari G<sup>29</sup>, Pereira DM<sup>52</sup>, Pérol O<sup>27</sup>,  
893 Pervez S<sup>21</sup>, Pomata A<sup>16</sup>, Popovic M<sup>28</sup>, Poveda A<sup>47</sup>, Prado CP<sup>6</sup>, Prager KM<sup>14</sup>, Ramieri  
894 G<sup>29</sup>, Rasul S<sup>3</sup>, Rego JN<sup>53</sup>, Reis RM<sup>18</sup>, Renard H<sup>4</sup>, Ricardi U<sup>36</sup>, Riva G<sup>29</sup>, Rodilla F<sup>2</sup>,  
895 Rodriguez I<sup>54</sup>, Rodríguez MI<sup>35</sup>, Ross A<sup>31</sup>, Roux P<sup>55</sup>, Saeed Ali T<sup>56</sup>, Saintigny P<sup>57</sup>,  
896 Santivañez J J<sup>26</sup>, Scapultampo-Neto C<sup>58</sup>, Segovia J<sup>17</sup>, Sena A<sup>10</sup>, Serrano R<sup>40</sup>, Sharma  
897 S<sup>38</sup>, Siefer O<sup>38</sup>, Smart S<sup>59</sup>, Sorroche BP<sup>18</sup>, Sosa C<sup>16</sup>, Souza JD<sup>60</sup>, Stura A<sup>15</sup>, Thomas  
898 S<sup>34</sup>, Torres O<sup>61</sup>, Tous S<sup>2</sup>, Ucross G<sup>12</sup>, Valenzuela A<sup>35</sup>, Vasconcelos de Podestá J<sup>10</sup>,  
899 Whitmarsh A<sup>34</sup>, Wright S<sup>62</sup>

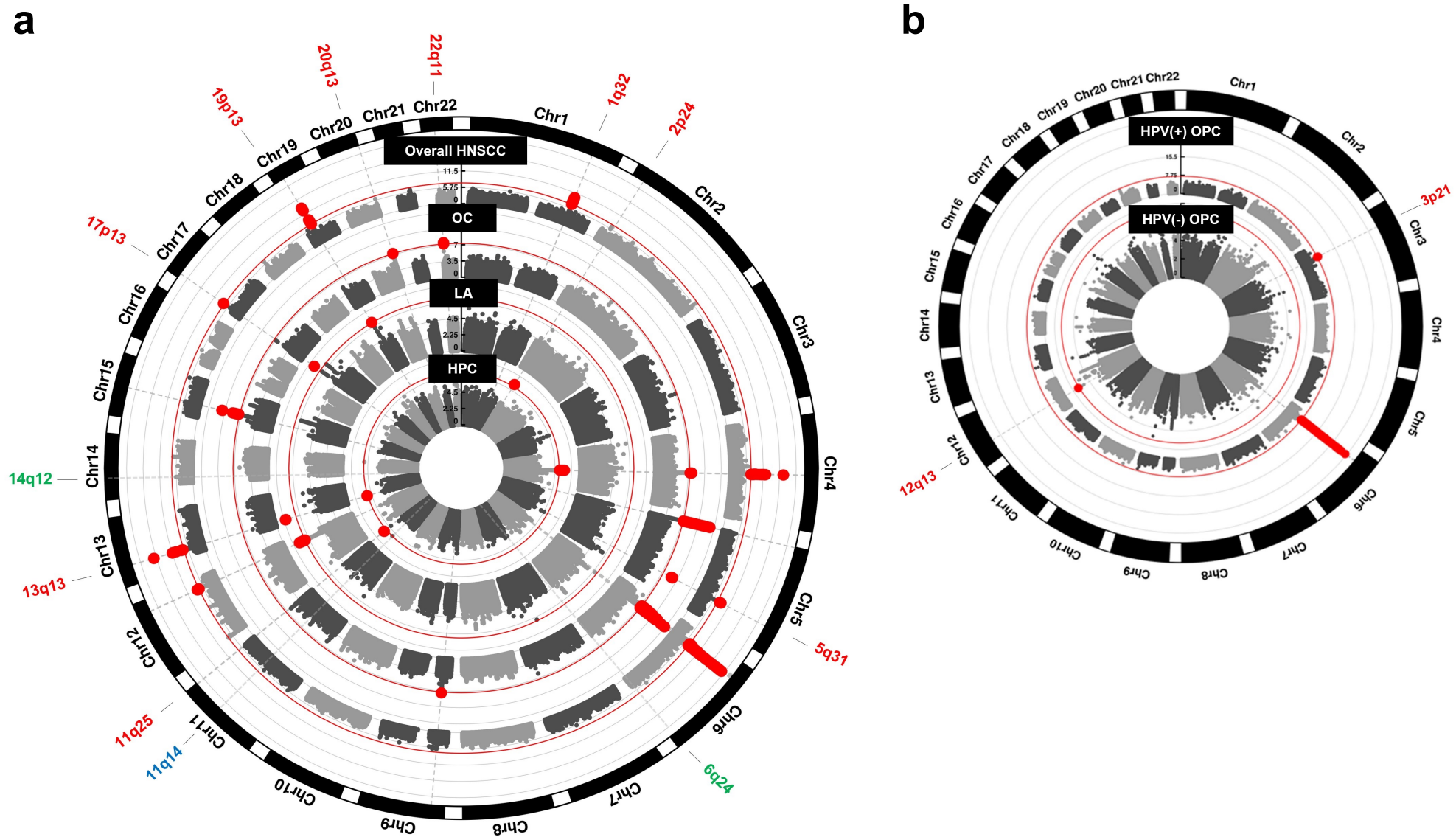
900

901 <sup>1</sup>H&N cancer Department, Universidad de Buenos Aires, Ciudad Autonoma de Buenos  
902 Aires, Argentina, <sup>2</sup>Catalan Institute of Oncology (ICO), Barcelona, Spain, <sup>3</sup>Department  
903 of Surgery, Dental Hygiene Program, Aga Khan University Hospital, Karachi, Pakistan,  
904 <sup>4</sup>Genomic Epidemiology Branch, International Agency for Research on cancer  
905 (IARC/WHO), Lyon, France, <sup>5</sup>Faculty of Science and Technology, University of Central  
906 Punjab, Lahore, Pakistan, <sup>6</sup>Group of Epidemiology and Statistics on Cancer, A.C.

907 Camargo Cancer Center, Sao Paulo, Brazil, <sup>7</sup>Servicio de Cabeza y Cuello, Instituto  
908 Nacional del Cáncer, Ministerio de Salud Pública y Bienestar Social, Capiatá,  
909 Paraguay, <sup>8</sup>Pathology and Laboratory Department, Fundación SantaFe de Bogotá,  
910 Bogotá, Colombia, <sup>9</sup>Postgraduate Program in Biotechnology, Universidade Federal do  
911 Espirito Santo, Vitoria, Brazil, <sup>10</sup>Head and Neck Surgery Division, Associação  
912 Feminina de Educação e Combate ao Câncer(AFECC), Hospital Santa Rita de Cássia,  
913 Vitoria, Brazil, <sup>11</sup>H&N cancer Department, Institute of Oncology Angel H. Roffo,  
914 University of Buenos Aires, Ciudad Autonoma de Buenos Aires, Argentina,  
915 <sup>12</sup>Department of Radiology, Division of Nuclear Medicine, Fundación SantaFe de  
916 Bogotá, Bogotá, Colombia, <sup>13</sup>Department of Pathology, Universidade Federal do  
917 Espirito Santo, Vitoria, Brazil, <sup>14</sup>Division of Infections and Cancer Epidemiology,  
918 German Cancer Research Center (DKFZ), Heidelberg, Germany, <sup>15</sup>Department of  
919 Medical Sciences, Cancer Epidemiology Unit, AOU Città della Salute e della Scienza  
920 di Torino, Turin, Italy, <sup>16</sup>Departamento de Anatomía Patológica, Instituto Nacional del  
921 Cáncer, Ministerio de Salud Pública y Bienestar Social, Capiatá, Paraguay,  
922 <sup>17</sup>Oncology Department, Fundación SantaFe de Bogotá, Bogotá, Colombia,  
923 <sup>18</sup>Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil,  
924 <sup>19</sup>Department of Head and Neck Surgery, Barretos Cancer Hospital, Barretos, Brazil,  
925 <sup>20</sup>Pathology Unit, AOU Città della Salute e della Scienza di Torino, Turin, Italy,  
926 <sup>21</sup>Department of Pathology and Laboratory Medicine, Section of Histopathology, Aga  
927 Khan University Hospital, Karachi, Pakistan, <sup>22</sup>INCA, Rio de Janeiro, Brazil, <sup>23</sup>School  
928 of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, United Kingdom,  
929 <sup>24</sup>Universidad Nacional de Cordoba, Cordoba, Argentina, <sup>25</sup>A.C Camargo Cancer  
930 Center, São Paulo, Brazil, <sup>26</sup>Department of Surgery, Head and Neck Division,  
931 Fundación SantaFe de Bogotá, Bogotá, Colombia, <sup>27</sup>Department Cancer  
932 Environnement, Centre Léon Bérard, Lyon, France, <sup>28</sup>Department of Medical  
933 Sciences, Cancer Epidemiology Unit, University of Turin, Turin, Italy, <sup>29</sup>Department of  
934 Surgical Sciences, University of Turin, Turin, Italy, <sup>30</sup>Metabolic Disorders Research  
935 Center, Golestan University of Medical Sciences, Gorgan, Iran, <sup>31</sup>School of Medicine,  
936 Dentistry and Nursing, University of Glasgow, Glasgow, United Kingdom,  
937 <sup>32</sup>Department of Pathology and Molecular Medicine, 2nd Faculty of Medicine, Charles  
938 University and Motol University Hospital, University Hospital in Motol, Prague, Czech  
939 Republic, <sup>33</sup>Department of Genetics, Genomics and Informatics, University of  
940 Tennessee Health Science Center, Memphis, USA, <sup>34</sup>Bristol Dental School, University

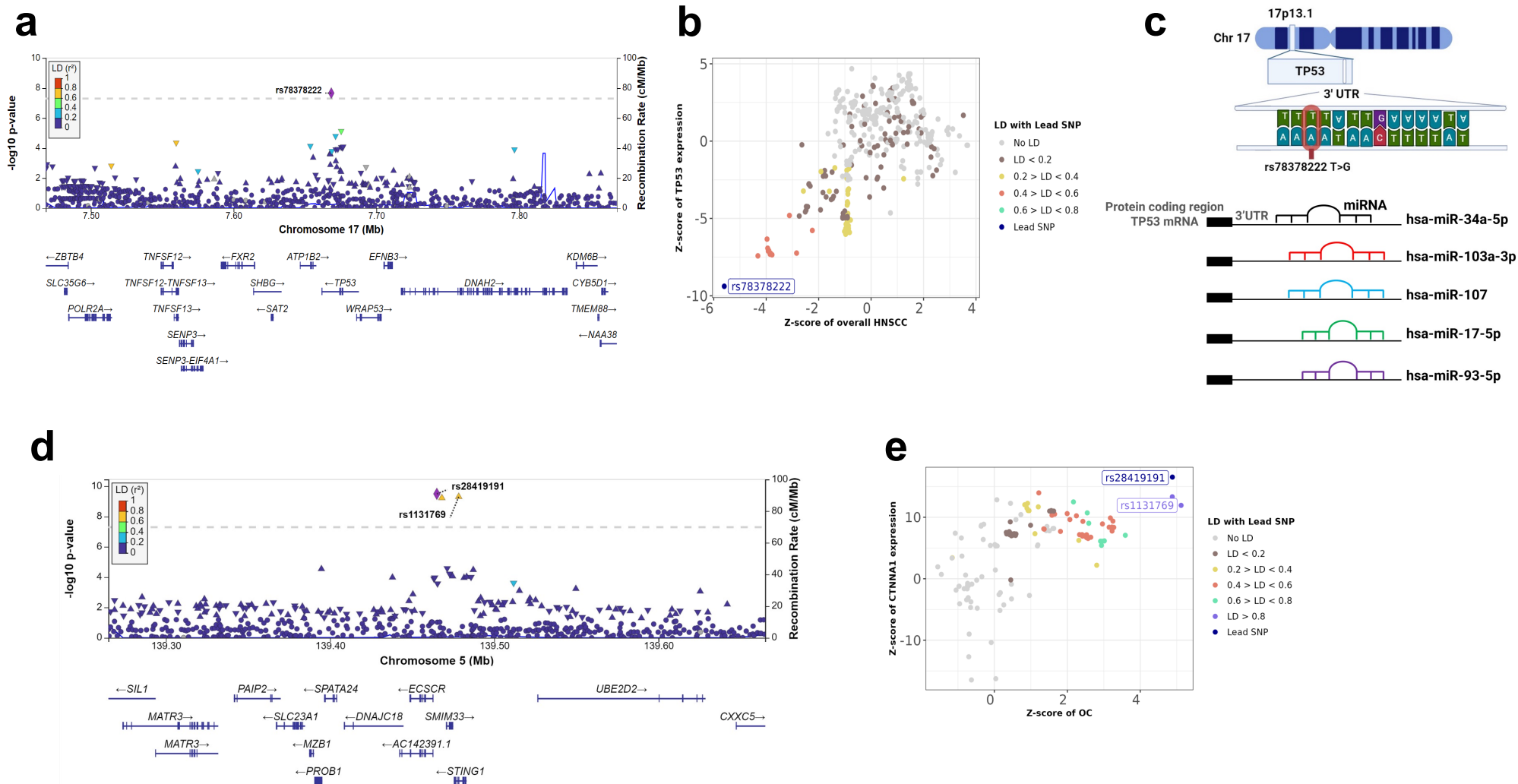
941 of Bristol, Bristol, United Kingdom, <sup>35</sup>Salud Pública, Instituto de Investigaiones en  
942 Ciencias de la Salud (IICS), Universidad Nacional de Asunción (UNA), San Lorenzo,  
943 Paraguay, <sup>36</sup>Department of Oncology, University of Turin, Turin, Italy, <sup>37</sup>Department of  
944 Surgery, Section of E.N.T, Aga Khan University Hospital, Karachi, Pakistan,  
945 <sup>38</sup>Department of Otorhinolaryngology, Head and Neck Surgery, University of Cologne,  
946 Cologne, Germany, <sup>39</sup>Department of Otorhinolaryngology and Head and Neck  
947 Surgery, University Hospital in Motol, Prague, Czech Republic, <sup>40</sup>Cátedra  
948 Otorrinonaringología, Hospital de Clínicas, Facultad de Ciencias Médicas,  
949 Universidad Nacional de Asunción, San Lorenzo, Paraguay, <sup>41</sup>Servicio de Oncología  
950 Clínica Hospital de Clínicas, Universidad de la República, Montevideo, Uruguay, <sup>42</sup>The  
951 Persian Gulf Tropical Medicine Research Center, The Persian Gulf Biomedical  
952 Sciences Research Institute, Bushehr University of Medical Sciences, Bushehr, Iran,  
953 <sup>43</sup>Statistics Genetic Lab, QIMR, Berghofer Medical Research Institute, Brisbane,  
954 Australia, <sup>44</sup>Department of Oral and Maxillofacial/Head and Neck Surgery, NHS  
955 Greater Glasgow and Clyde, Glasgow, United Kingdom, <sup>45</sup>Hospital Câncer Araújo  
956 Jorge, Goiânia, Brazil, <sup>46</sup>Hospital Câncer Araújo jorge, Goiânia, Brazil, <sup>47</sup>Faculty of  
957 Medicine, El Bosque University, Bogotá, Colombia, <sup>48</sup>Oncology Department, Division  
958 of Radiotherapy, Fundación SantaFe de Bogotá, Bogotá, Colombia, <sup>49</sup>Goiânia Cancer  
959 Registry (BR), Goiânia, Brazil, <sup>50</sup>Department of Oncology, Division of Medical  
960 Oncology, AOU Città della Salute e della Scienza di Torino, Turin, Italy, <sup>51</sup>Beatson West  
961 of Scotland Cancer Centre, NHS Greater Glasgow and Clyde, Glasgow, United  
962 Kingdom, <sup>52</sup>Radiation Oncology Department, Institute of Oncology Angel H. Roffo,  
963 University of Buenos Aires, Ciudad Autonoma de Buenos Aires, Argentina, <sup>53</sup>Clinical  
964 Research Center, Associação Feminina de Educação e Combate ao Câncer(AFECC),  
965 Hospital Santa Rita de Cássia, Vitória, Brazil, <sup>54</sup>Laboratorio de Anatomía Patológica,  
966 Hospital de Clínicas, Facultad de Ciencias Médicas, Universidad Nacional de  
967 Asunción, San Lorenzo, Paraguay, <sup>55</sup>Department of Surgery, Centre Léon Bérard,  
968 Lyon, France, <sup>56</sup>School of Nursing and Midwifery, Aga Khan University Hospital,  
969 Karachi, Pakistan, <sup>57</sup>Centre Léon Bérard, Lyon, France, <sup>58</sup>Pathology and Molecular  
970 Diagnostics Service, Barretos Cancer Hospital, Barretos, Brazil, <sup>59</sup>University of  
971 Glasgow, Glasgow, United Kingdom, <sup>60</sup>Epidemiology and Statistics Group, Research  
972 Center, A.C Camargo Cancer Center, São Paulo, Brazil, <sup>61</sup>Radiology Department,  
973 Fundación SantaFe de Bogotá, Bogotá, Colombia, <sup>62</sup>Institute of Cancer Sciences,  
974 University of Glasgow, Glasgow, United Kingdom.



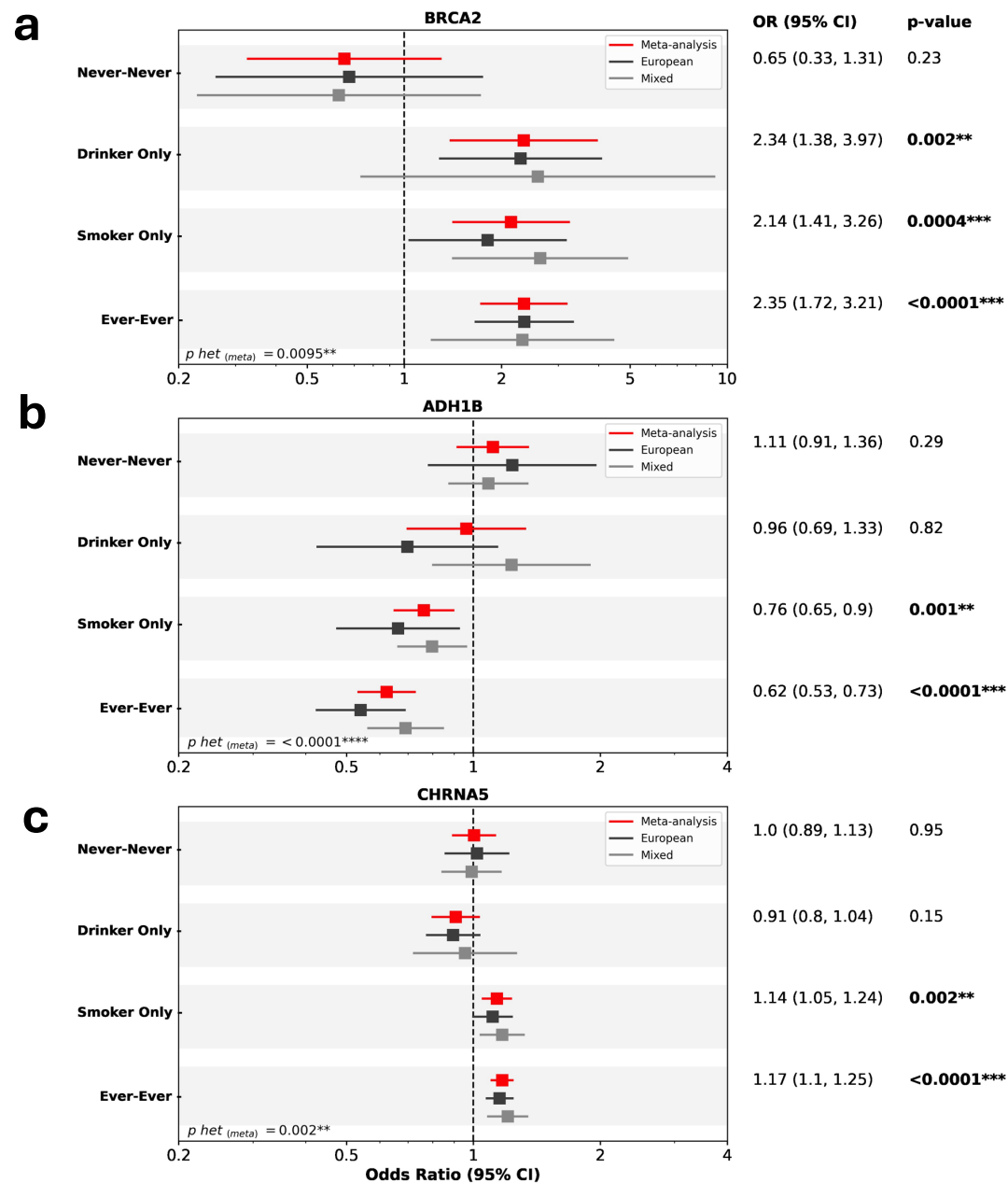


**Figure 1. Novel loci identified for HNSCC.** a) Circular Manhattan plots showing novel risk loci identified in this study. Red labels indicate the cytogenetic locations of novel signals identified in meta-analyses for all sites combined or subsite-specific. Blue labels represent novel loci identified in the European group only, and green labels indicate novel loci identified in the Mixed group only. Red lines mark the threshold for genome-wide significance ( $p = 5 \times 10^{-8}$ ). b) Circular Manhattan plots from the GWAS analyses of HPV(+) and HPV(-) oropharyngeal cancer. Separate Manhattan plots for each group can be found in Supplementary Figure 1.

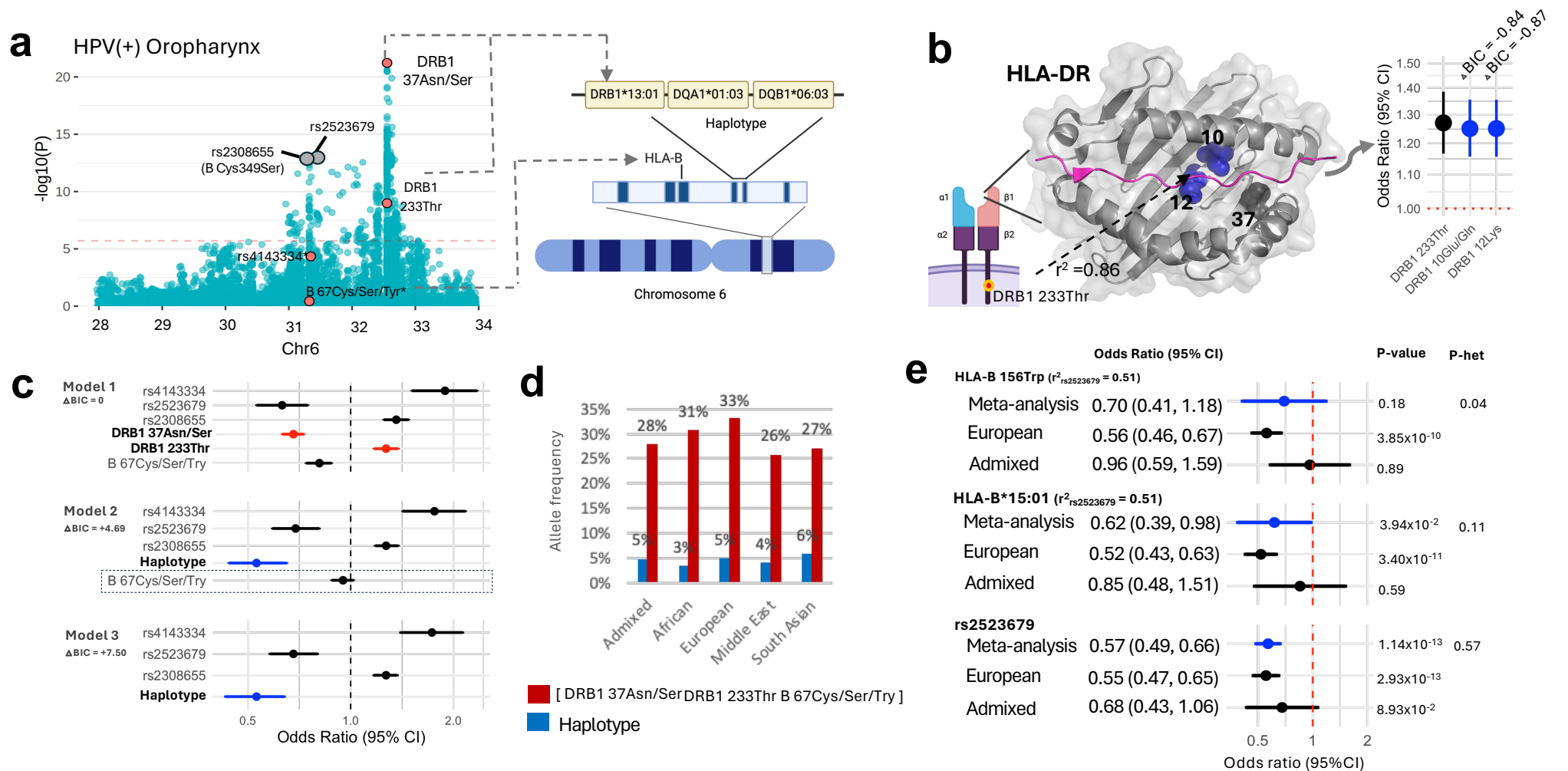




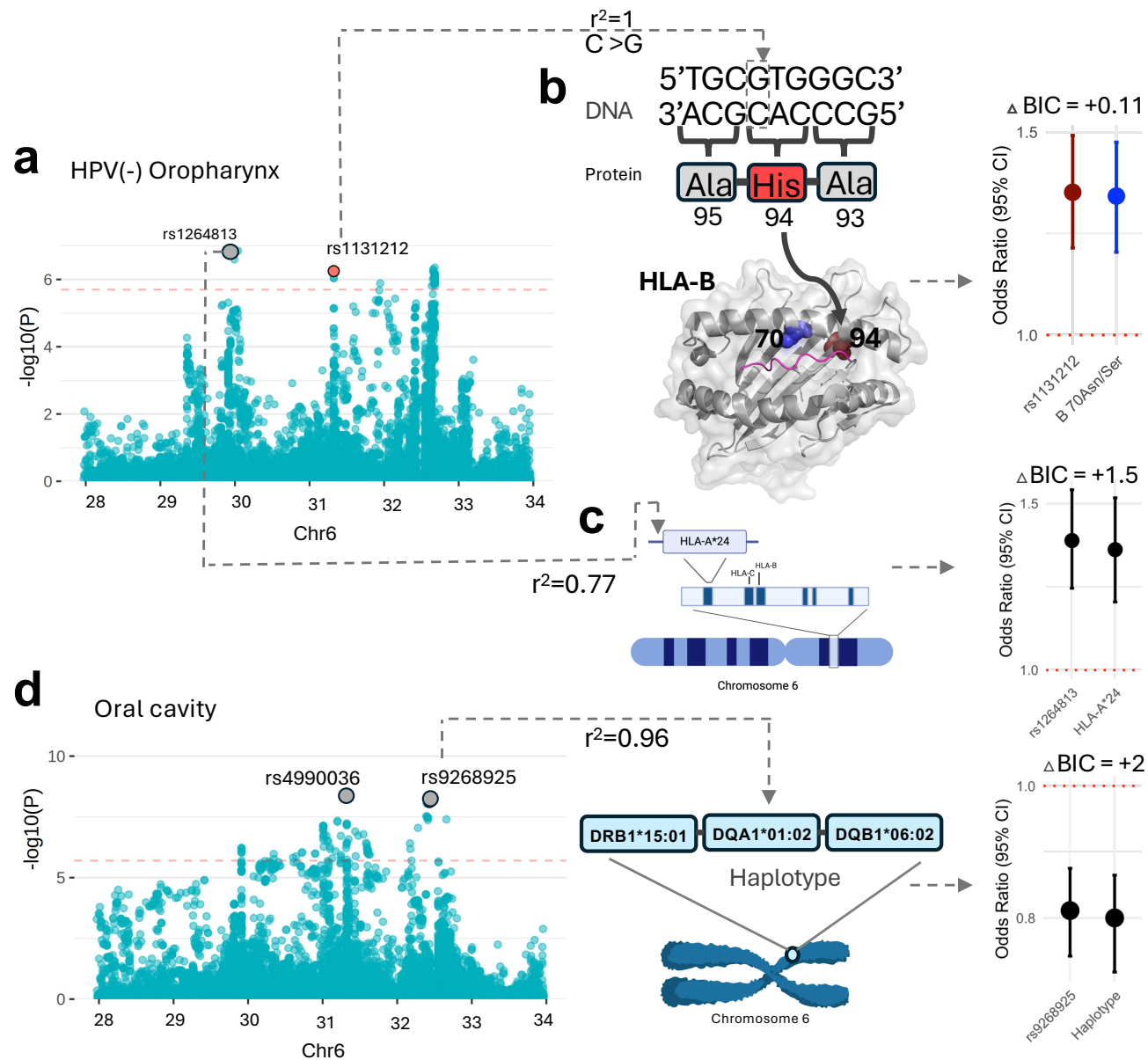
**Figure 2. Regional and Colocalization Analysis of *TP53* and *STING1* Variants.** a) Regional plot of the 3' UTR variant rs78378222 in *TP53* at chromosome 17p13. The x-axis represents chromosomal location, while the y-axis displays the  $-\log_{10}$  p-value. The dotted line marks the genome-wide significance threshold of  $5 \times 10^{-8}$ . Single nucleotide polymorphisms (SNPs) are color-coded according to their linkage disequilibrium ( $r^2$ ) with rs78378222. b) Z-Z locus plot showing rs78378222, the lead variant, is associated with reduced *TP53* expression in whole blood, with a high PP4 score of 99%. c) The cytogenetic location of rs78378222, along with its sequence and allele change, is shown at the chromosomal level. The variant overlaps with multiple predicted microRNA binding sites. d) Regional plot of rs28419191 (intergenic) and rs1131769 (*STING1*) at 5q31 from the oral cavity (OC) meta-analysis ( $r^2 = 0.93$ ). e) Z-Z locus plot showing colocalization of rs28419191 and rs1131769 with *CTNNA1* expression in whole blood, both with a PP4 score of 99%.



**Figure 3. Gene-environment interactions with alcohol and smoking.** Effect estimates for a) rs11571833 (*BRCA2*) b) rs1229984 (*ADH1B*) and c) rs58365910 (*CHRNA5*) stratified by smoking and drinking (Never smoker-Never drinker, Smoker Only, Drinker Only, and Ever smoker-Ever drinker) from the meta-analysis, and within European, and Mixed groups. Only the odds ratios, p-value and p-heterogeneity ( $p_{het}$ ) for the meta-analysis are shown here.



**Figure 4. Cross-ancestry HLA risk loci of HPV(+) OPC.** a) Manhattan plots showing all independent lead variants for risk of HPV(+) OPC. Variants highlighted under significance threshold reached significance in later rounds; only the plot from the first round of stepwise analysis is shown here. Novel variants are orange; known variants are grey. The horizontal red line reflects the HLA significance threshold ( $p < 2.4 \times 10^{-6}$ ). DRB1 37Asn/Ser, DRB1 233Thr, are within DRB1\*13:01-DQA1\*01:03-DQB1\*06:03 while HLA-B67Cys/Ser/Try was associated with the haplotype. b) Out of the five interchangeable amino acid residues in LD with DRB1 233Thr with  $\Delta BIC \pm 2$ , DRB1 12Lys and DRB1 10Glu/Gln are in the HLA-DR binding pocket. c) Model accuracy and risk estimates of amino acid residues and haplotypes. Model 1: identified from fine-mapping, used as the baseline reference model; Model 2: replaces DRB1 37Asn/Ser, DRB1 233Thr with the haplotype, effect of HLA-B 67Cys/Ser/Try disappears; Model 3: All 3 amino acids replaced with haplotype. d) Allele frequencies of DRB1\*13:01-DQA1\*01:03-DQB1\*06:03 and of having all three amino acid residues by ancestry. e) The HLA-B 156Trp amino acid change and the HLA-B 15:01 allele are specific to European ancestry, but rs2523679 variant which is in LD with both, has a cross-ancestral effect.



**Figure 5. Novel HLA risk loci for HPV(-) oropharynx and oral cavity cancer.** Manhattan plots display all independent lead variants of risk for each HNSCC subsite. Novel variants are highlighted in red; known variants are in grey. The horizontal red line reflects the HLA significance threshold ( $p < 2.4 \times 10^{-6}$ ) a) HPV(-) oropharynx: The lead SNP, b) rs1131212, causes an amino acid change from Gln to His at residue 94 located in the HLA-B protein binding pocket (PDB ID: 2BVP). This variant is in LD ( $r^2=1$ ) with 70Asn/Ser. The right panel shows the comparable risk effects of the two related signals. The known SNP, c) rs1264813, is in high LD ( $r^2=0.77$ ) with HLA-A\*24 allele and shows comparable risk effects shown in right panel. d) Oral cavity: The lead SNP, rs9268925, is highly correlated with a novel risk haplotype, DRB1\*15:01-DQA1\*01:02-DQB1\*06:02, and has a similar risk effect, as shown in the right panel. Model accuracy difference ( $\Delta BIC$ ) between original model in presence of all independent lead variants and the model replacing the lead variant with related amino acid residue, allele or haplotype, lower than 2 confer equivalent risk.

**Table 1.** Summary of Novel Genetic Variants Identified in European and Mixed Groups Through GWAS and Meta-Analysis

Population	Subsite	rsID	Mapped/Nearest Gene	CHR	BP (GRCh38)	Cytogenetic Position	Major Allele	Effect Allele	EAF	OR (95%CI)	p-value
Meta-Analysis	All sites combined	rs61817953	<i>PIK3C2B</i>	1	204493484	1q32	G	A	-	0.90 (0.87, 0.93)	1.17x10 <sup>-8</sup>
		rs6679311	<i>MDM4</i>	1	204590548	1q32	C	T	-	1.11 (1.07, 1.14)	1.25x10 <sup>-10</sup>
		rs1131769	<i>STING1</i>	5	139478334	5q31	C	T	-	1.13 (1.09, 1.18)	2.38x09 <sup>-09</sup>
		rs7334543	<i>BRC42</i>	13	32399139	13q13	A	G	-	0.91 (0.88, 0.94)	2.39x08 <sup>-08</sup>
		rs78378222 <sup>a</sup>	<i>TP53</i>	17	7668434	17p13	T	G	-	0.62 (0.52, 0.73)	2.16x08 <sup>-08</sup>
	rs10419397	<i>ANKLE1</i>	19	17280519	19p13	G	A	-	1.13 (1.10, 1.17)	1.21x14 <sup>-14</sup>	
	OC	rs28419191	<i>ECSCR</i>	5	139465014	5q31	C	T	-	1.23 (1.15, 1.31)	3.16x10 <sup>-10</sup>
		rs67351073	<i>ZGPAT</i>	20	63704213	20q13	GA	G	-	0.78 (0.72, 0.85)	4.45x08 <sup>-08</sup>
		rs577454702 <sup>a</sup>	<i>MAPK1</i>	22	21778123	22q11	A	C	-	2.60 (1.86, 3.65)	2.53x08 <sup>-08</sup>
	LA	rs55831773	<i>ATP1B2</i>	17	7655719	17p13	C	T	-	1.21 (1.13, 1.29)	5.1x09 <sup>-09</sup>
		rs10419397	<i>ANKLE1</i>	19	17280519	19p13	G	A	-	1.18 (1.10, 1.26)	4.33x08 <sup>-08</sup>
	HPC	rs138707495	<i>GDF7</i>	2	20677150	2p24	T	TA	-	3.06 (2.07, 4.53)	2.33x08 <sup>-08</sup>
		rs77750788	<i>IGSF9B</i>	11	133936692	11q25	G	A	-	2.07 (1.61, 2.68)	2.03x08 <sup>-08</sup>
rs181194133 <sup>a</sup>		<i>OPCML</i>	11	132728232	11q25	G	A	-	3.44 (2.24, 5.31)	2.09x08 <sup>-08</sup>	
European Ancestry GWAS	HPC	rs181777026	<i>TENM4</i>	11	81037815	11q14	C	T	0.012	2.81 (1.94, 4.05)	3.78x08 <sup>-08</sup>
	HPV(-) OPC	rs112726671	<i>VDR</i>	12	47926100	12q13	A	G	0.016	2.28 (1.70, 3.07)	4.03x08 <sup>-08</sup>
	HPV(+) OPC	rs1520483	<i>LTF</i>	3	46468719	3p21	C	T	0.400	1.23 (1.14, 1.32)	2.19x08 <sup>-08</sup>
Mixed Ancestry Group GWAS	LA	rs200410709	<i>STXBP6</i>	14	25417834	14q12	CT	C	0.013	3.38 (2.26, 5.07)	3.57x09 <sup>-09</sup>
	HPC	rs150899739	<i>SASH1</i>	6	148061934	6q24	G	A	0.008	5.84 (3.17, 10.76)	1.47x08 <sup>-08</sup>

Novel variants identified in meta-analysis and by group across subsites. Full list of significant variants can be found in Supplementary Table 3.

CHR = Chromosome; BP = Base-pair position; EAF = Effect allele frequency; OR = Odds ratio; OC = Oral cavity; LA = Larynx; HPC = Hypopharynx; OPC = Oropharynx.

<sup>a</sup> These variants were only identified in European GWAS and were not present in the mixed group, as their minor allele frequency was below the 0.05% threshold determined in our analyses.

**Table 2.** Summary of Novel Genetic Variants Identified Across Ancestry-Specific and Meta-Analysis of HLA-fine mapping in all sites combined and subsite specific

Population	Subsite	Variant	Gene	Position (hg19)	Locus; Cytoband	Impact	Ref <sup>a</sup>	Effect Allele <sup>a</sup>	OR (95% CI)	p-value <sup>b</sup>
Meta-Analysis	All sites combined	Chr6:33046667	HLA-DPB1	Chr6:33046667	class II	intron	C	T	1.11 (1.07,1.14)	1.32x10 <sup>-8</sup>
		rs28360051	PSORS1C3	Chr6:31142261		intron	G	A	1.23 (1.14,1.34)	1.91x10 <sup>-7</sup>
	HPV(-) OPC	rs1131212	HLA-B	Chr6:31324526	class I	Gln94His	C	G	1.33 (1.19,1.49)	5.33x10 <sup>-7</sup>
	HPV(+) OPC	DRB1 37Asn/Ser	HLA-DRB1	Chr6:32552051	class II	Amino acid change	Ab	Pr	0.68 (0.63,0.73)	3.22x10 <sup>-23</sup>
		rs4143334	ZDHHC20P2	Chr6:31348200	class I	Non coding transcript exon	A	G	1.89 (1.51,2.35)	1.91x10 <sup>-8</sup>
		DRB1 233Thr	HLA-DRB1	Chr6:32548048	class II	Amino acid change	Ab	Pr	1.27 (1.17,1.38)	7.15x10 <sup>-9</sup>
B 67Cys/Ser/Tyr		HLA-B	Chr6:31324536	class I	Amino acid change	Ab	Pr	0.81 (0.74,0.88)	1.33x10 <sup>-06</sup>	
Admixed	All sites combined	rs1536036	ITPR3	Chr6:33632014		Intron	A	G	0.85 (0.80,0.91)	8.42x10 <sup>-7</sup>
European	OC	DRB1 74Ala/Leu/Del	HLA-DRB1	Chr6:32552625	class II	Amino acid change	Ab	Pr	0.82 (0.77,0.87)	4.94x10 <sup>-10</sup>
		rs9267280	MICB-DT	Chr6:31457633	class I	Intron	G	A	1.32 (1.19,1.47)	3.48x10 <sup>-7</sup>
	HPV(+) OPC	HLA-B*51:01	HLA-B	Chr6:31321767	class I	Amino acid change	Ab	Pr	1.9 (1.55,2.31)	3.6x10 <sup>-10</sup>

Novel variants identified in meta-analysis and ancestry specific groups shown here. No novel variants were identified within Middle Eastern, African and South Asian populations. Full list of variants in this region can be found in Supplementary Table 10. All variants were tested for independence, which was defined by linkage disequilibrium ( $R^2$ ) < 0.3 and Bonferroni threshold of  $P > 10^{-6}$  when conditioning on variants from other subsites or with previously identified variants. Further details can be found in Supplementary Tables 8 and 9

<sup>a</sup> Ref/A1 allele is in binary marker format (Ab = Absent, Pr = Present) of classical HLA alleles, amino acid residue, HLA intragenic, insertion/deletions. (see: <https://imputationserver.readthedocs.io/en/latest/pipeline>)

<sup>b</sup> meta-analysis Pvalue of the final model including all significant independent variants adjusted by sex, imputation batch and PCs

HLA significance level =  $2.4 \times 10^{-6}$  considering all variants in chr6