Somatic alterations in *TP53*, *PIK3CA*, and *CCND1* reveal exposure-driven patterns in oral and oropharyngeal tumors

Ricardo Cortez Cardoso Penha[1], PhD, Shama Virani[1*], PhD, Jeremiah R Holt[2,3*], PhD, Xiaobei Zhao[3*], PhD, Paul Little[4], PhD, Myriam Brossard[5], PhD, Valérie Gaborieau[1], Degree: N/A., Maja Milojevic[6], PhD, Laura Torrens[1], PhD, Hanla A Park[1], PhD, Tom Dudding[7], PhD, Wendell G Yarbrough[8], MD, MMHC, FACS, Miranda Pring[7], PhD, Vonn Walter[9], PhD, Scott V Bratman[10], PhD, Wolfgang Ahrens[11], PhD, Laia Alemany[12], PhD, Cristina Canova[13], PhD, David I Conway[14], PhD, Claire M Healy[15], PhD, Ivana Holcatova[16], PhD, Luis Felipe Ribeiro Pinto[17], PhD, Maria Paula Curado[18], PhD, Sandra Ventorin von Zeidler[19], PhD, Jose Carlos de Oliviera[20], PhD, José Roberto Vasconcelos de Podestá[21], MD, Marta Vilensky[22], PhD, Areti Lagiou[23], PhD, Gary J Macfarlane[24], DSc, Jerry Polesel[25], MSc, Lorenzo Richiardi[26], PhD, Max Robinson[27], PhD, Ariana Znaor[28], PhD, Jose Zevallos[29], MD, MPH, Sandra Perdomo[1], PhD, Paul Brennan[1], PhD, Rayjean J Hung[5], PhD, Tim Waterboer[30], PhD, Geoffrey Liu[31], MD, MSc, Andrew F Olshan[32], PhD, Brenda Diergaarde[33#], PhD, D. Neil Hayes[3#], MD, MS, James D McKay[1#], PhD

* Equally contributing authors; # Equally contributing senior authors

[1]Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France, [2]Medicine, Hematology and Oncology, Center for Cancer Research, Memphis, TN USA, [3]University of Tennessee Health Science Center, Center for Cancer Research Memphis, TN 38103 USA, [4]Center for Cancer Research, University of Tennessee Health Science Center, Memphis, TN 38103 USA, [5]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada,[6]Centre of Excellence for Health, Immunity and Infections, Rigshospitalet, Copenhagen University Hospital, Copenhagen, 2100 Denmark, [7]Bristol Dental School, Bristol University, Bristol, UK, [8]Department of Otolaryngology/Head and Neck Surgery, UNC School of Medicine, University of North Carolina, Chapel Hill, NC USA, [9]Department of Public Health Sciences, Penn State University College of Medicine, Hershey, PA 17033 USA, [10]Department of Radiation Oncology, Princess Margaret Cancer Centre, University of Toronto, Toronto, Canada, [11]Epidemiological Methods and Etiological Research, Leibniz Institute for Prevention Research and Epidemiology-BIPS, Bremen, Bremen 28359 Germany, [12]Catalan Institute of Oncology, L'Hospitalet, Barcelona, Spain, [13]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thoraco-Vascular Sciences and Public Health, University of Padova, Padova, 35100 Italy, [14]School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, UK, [15]School of Dental Science, Dublin Dental University Hospital, Trinity College Dublin, Dublin, Ireland, [16]Institute of Hygiene & Epidemiology, 1st Faculty of Medicine, Charles University; and Department of Oncology, University Hospital Motol & 2nd Faculty of Medicine, Charles University, Prague, Czech Republic, [17]Programa de Carcinogênese Molecular, Instituto Nacional de Câncer - INCA, Rio de Janeiro, Brazil, [18]Epidemiology and Statistics Group, Research Center, A.C Camargo Cancer Center, São Paulo, Brazil, [19]Pathology Department, Federal University of Espírito Santo, Vitória, Brazil, [20]Araújo Jorge Cancer Hospital, Associação de Combate ao Câncer em Goiás, Goiania, Brazil, [21]Associação Feminina de Educação e Combate ao Câncer - Hospital Santa Rita de Cássia, Vitória, Brazil, [22]Instituto de Oncologia Angel H Roffo, Universidad de Buenos Aires, Buenos Aires, Argentina, [23]Department of Public and Community Health, School of Public Health, University of West Attica, Athens, Greece, [24]Epidemiology Group, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB25 2ZD UK, [25]Unit of Cancer Epidemiology, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Aviano, 33081 Italy, [26]Department of Medical Sciences, University of Turin, Turin, 10127 Italy, [27]Cellular Pathology, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, Tyne and Wear NE1 4LP UK, [28]Cancer Surveillance Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France, [29]Department of Otolaryngology, University of Pittsburgh Medical Center, Pittsburgh, PA USA, [30]Infections and Cancer Epidemiology, German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), Heidelberg, 69120 Germany, [31]Medicine, Epidemiology, Medical Oncology, Princess Margaret Cancer Centre, University of Toronto, Toronto, ON M5G2M9 Canada, [32]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, USA,

© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

[33]Department of Human Genetics, School of Public Health, University of Pittsburgh, and UPMC Hillman Cancer Center, Pittsburgh, PA USA

Corresponding author: James D McKay (email: mckayj@iarc.who.int)

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Abstract

Tobacco, alcohol, and human papillomavirus (HPV) are major risk factors for oral cavity (OC) and oropharyngeal (OPC) cancers, yet their specific contributions to the tumor molecular landscape remain incompletely understood. We analyzed somatic alterations in 1,086 tumors (606 OC, 480 OPC, including 287 HPV(+)OPC) profiled with a 1,109-gene panel and evaluated their associations with exposures. *PIK3CA* alterations were widespread but context-specific: HPV(+)OPC featured APOBEC-driven mutations and co-occurrence of *FGFR3* mutations in a subset of cases, while HPV(-)OPC and OC were marked by 3q26/28 (*PIK3CA*/*TP63*/*SOX2*) amplifications. *TP53* was the most frequently mutated gene in HPV(-) tumors, with strong enrichment among smokers in HPV(-)OPC, whereas *TP53* mutation frequencies in OC were consistently high regardless of smoking status. *CCND1* amplifications were largely restricted to HPV(-) tumors and were most frequent in individuals reporting combined tobacco/alcohol use. This combined effect was replicated in the TCGA OC-OPC cohort and mirrored at the transcriptomic level, where *CCND1* mRNA expression was highest in dual users. Within OC, characterization of one of the largest cohorts of tumors from never smoker-never drinkers to date (N=100) showed they were more likely to be from female patients and to exhibit enrichment for *FAT1* and *CASP8* mutations compared to tumors from ever smoker-ever drinkers (N=353). Together, these findings define distinct, exposure-driven patterns of somatic alterations by subsite and HPV status, and provide a molecular framework that may inform etiologic-based risk stratification in head and neck cancer.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Introduction

Oral cavity (OC) and oropharyngeal (OPC) tumors are types of head and neck squamous cell carcinoma (HNSCC) affecting the mouth and pharynx, respectively[1]. In 2022, an estimated 500,000 new cases of OC and OPC were diagnosed worldwide, accounting for approximately 2.5% of the total cancer burden[2]. Tobacco smoking and alcohol consumption are well-established risk factors for both cancer types, with a multiplicative effect on risk among individuals exposed to both[2,3]. Persistent infection with human papillomavirus (HPV), particularly the high-risk subtype HPV16, has also emerged as a major risk factor in OPC[1,2,4]. HPV-driven HPV(+)OPC is now considered a distinct disease entity due to its unique etiology, molecular characteristics, and favorable prognosis compared to non-HPV-driven HPV(-)OPC and OC[4–7].

Previous genomic studies have identified both shared and distinct somatic alterations in OC and OPC, often linked to HPV, tobacco, or alcohol exposure. In HPV(+)OPC, carcinogenesis is driven by the viral oncoproteins E6 and E7, which degrade *TP53* and *RB1* and promote uncontrolled cell-cycle progression[7–10]. APOBEC-mediated mutagenesis is also more prevalent in HPV(+)OPC, contributing to hotspot mutations in *PIK3CA* and *FGFR3*[7–10]. Frequently observed copy number alterations (CNAs) include amplifications of 3q26/28 (*PIK3CA*, *TP63*, *SOX2*) and 20q11 (*E2F1*), and deletions in 14q32 (*TRAF3*)[7,9–11]. By contrast, HPV(-) tumors commonly harbor loss-of-function mutations in *TP53* and inactivation of *CDKN2A* (encoding p16INK4a), alongside a higher CNA burden indicative of greater genomic instability[7,9]. In HPV(-) tumors, *TP53*[12], *NSD1*, and *COL1A11*[13] mutations are enriched among smokers. Amplifications of 11q13 (including *CCND1*, *FADD*, and *CTTN*), particularly *CCND1* amplification and overexpression, are also more common in smokers[14,15]. *CASP8* mutations, in contrast, are enriched in copy-neutral OC tumors from never-users of tobacco and alcohol[9,12,13,16].

Improved knowledge of the somatic alteration patterns in OC and OPC and their relationship with modifiable risk factors such as tobacco and alcohol use are essential for advancing our understanding of head and neck carcinogenesis, and may help inform prevention strategies and guide the development of therapeutic targets. To address this, we performed targeted sequencing (1,109-gene panel) on well-annotated tumor specimens from a large cohort of OC and OPC patients from North America, South America, and Europe. Our dataset includes detailed clinical and demographic data, including information on HPV status, tobacco and

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

alcohol use, alongside sequencing results for 606 OC and 480 OPC tumors, including 287 HPV(+)OPC cases. This represents one of the largest datasets of HPV(+)OPC tumors sequenced to date. Previous studies have been limited by smaller sample sizes[9,12,17], particularly for HPV(+)OPC, and lacked statistical power to assess risk-factor associations across subsites. Here, we report subsite-specific somatic alterations in frequently altered genes, their relationships with tobacco and alcohol exposures (alone and in combination), and their distribution by HPV status. These analyses provide new insights into the molecular etiology of OC and OPC and the relationship with environmental and viral risk factors.

Methods

Study population and demographic and clinical data

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. Informed consent was obtained for all study participants. Ethical approvals were obtained from Local and Federal Research Ethics Committees, and from the IARC Ethics Committee (IEC Project 17-10A4).

The study population consists of OC and OPC patients with tumor tissue available for sequencing from the VOYAGER (human papillomaVirus, Oral and oropharYngeal cAncer GEnomic Research) consortium[18] and an additional study from South America (INTERCHANGE)[19]. VOYAGER includes five studies: the Alcohol-related cancers and genetic susceptibility in Europe study (ARCAGE; Europe)[20], the Head and Neck 5000 study (HN5000; United Kingdom)[21], the Toronto Mount Sinai Hospital-Princess Margaret (MSH-PMH) study (Toronto; Canada)[22], the University of Pittsburgh case-control study on head and neck cancer (Pittsburgh; US)[23], and the Carolina Head and Neck Cancer Epidemiology study (CHANCE; US)[24]. Demographic data, information on established risk factors, such as tobacco and alcohol use history, and clinical variables were collected by all six studies, shared, and harmonized. Available clinical variables include age at diagnosis, ICD-10 code, tumor, nodal and metastasis (TNM) status, stage (AJCC 7th edition), and tumor p16 and/or HPV DNA status. The current study includes pathologically confirmed OC and OPC patients diagnosed between 2002-2017. Oral cavity cases include ICD-10 codes C00.3-C00.6, C00.8-C00.9, C02.0-C02.3, C02.8, C02.9, C03.0-C03.9, C04.0-C04.9, C05.0, C05.8, C05.9, and C06.0-C06.9; oropharynx cases include C01, C01.9, C02.4, C05.1, C05.2, and C09.0-C10.9. All tumor samples included in this study were primary, treatment-naïve squamous cell carcinomas.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

HPV16 E6 serology data was available for 90% of the OPC cases included and was prioritized to assign HPV status[25]. If HPV16 E6 serology data was not available, HPV status was assigned based on concordant results between p16 immunohistochemistry (IHC) and high-risk HPV DNA *in situ* hybridization (ISH)[26] or HPV DNA sequencing if their sequencing results showed the presence of high-quality HPV16 sequence reads (average GC content between 34-45%, average insert size between 140-220, genome coverage > 85%, and median read depth > 450). If p16 IHC and the second marker were discordant or only one marker was available, then HPV status was set to "unknown".

Tobacco smoking and alcohol consumption information were self-reported and collected using study specific questionnaires at time of recruitment into the studies[19–24]. Information available on tobacco smoking included smoking status, number of cigarettes per day and pack-years. An ever smoker was defined as an individual who smoked at least one cigarette per day for one year or longer. Ever smokers were classified as heavy smokers if they smoked ≥20 pack-years and as light-to-moderate smokers if they smoked <20 pack-years (consistent with U.S. Preventive Services Task Force criteria[27]). Information available on alcohol consumption included alcohol drinking status and number of drinks per day (drinking intensity). We classified those who reported no alcohol consumption as never drinkers, and those that did as ever drinkers. For HN5000 study participants, information was available on current drinking status only, and those who were not current drinkers (9% of the HN5000 study population) were classified as never drinkers. Ever drinkers were classified as heavy drinkers if they drank >2 drinks per day (for males) and >1 drinks per day (for females), and as light-to-moderate drinkers if they drank ≤2 drinks per day (for males) and ≤1 drinks per day (for females), according to the current guidelines of the Centers for Disease Control and Prevention (CDC)[28].

Library preparation, sequencing, alignment and data processing

DNA was extracted using Qiagen DNA extraction kits as recommended by the manufacturer. A scalable Next-Generation Sequencing (NGS) assay (UTseq) using the Agilent SureSelect protocol was developed to analyze DNA extracted from tumor specimens, following the manufacturer's specifications. This assay was implemented at the University of Tennessee Health Science Center for Cancer Research (Memphis, TN, US). UTseq uses the ClinSeq gene panel (version 10)[29] (Source data), which includes 1,109 cancer-associated genes and clinically relevant targets, along with non-human sequences captured by the bait set, such as full-length viral genomes for human papillomavirus (HPV) strains 16 and 18 (PAVE reference numbers:

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

gi|333031 and gi|60975, respectively)[30]. Libraries containing paired-end reads of 150 bases in length were sequenced on Illumina sequencers (NextSeq, HiSeq, or NovaSeq), following standard NGS protocols[31], with a mean target depth coverage of 500x.

FASTQ format files were generated for each prepared library and processed through the UTseq pipeline for quality control (QC) assessment. The sequencing reads were aligned to reference databases that included the human genome [hs37d5 (GRCh37/hg19)] and 27 viral genomes, including 19 HPV strains (HPV16, HPV18, HPV26, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV53, HPV56, HPV58, HPV59, HPV66, HPV68, HPV70, HPV73, and HPV82) curated in the PAVE database using the Burrows-Wheeler Aligner[32] (BWA; version 0.7.9a). After alignment, BAM files were filtered for duplicate reads and sorted using the bamsort command from biobambam[33] (version 0.0.191). To refine alignments and enhance structural variant detection in both human and viral genomes, Assembly-Based ReAligner (ABRA)[34] (version 0.96) was used for de novo assembly of target reads followed by realignment.

UNMASC tumor-only variant calling and QC

Tumor-only variant calling was performed using the Unmatched Normals and Mutant Allele Status Characterization (UNMASC) tool[29] (https://github.com/pllittle/UNMASC). Briefly, tumor-only variant calling was performed by pooling single-nucleotide variants and indels called by Strelka[35] (version 2.9.10) with the GRCh37 reference genome. To ensure high-confidence somatic variant calls, we applied a rigorous QC pipeline to the variants identified by UNMASC[29]. Briefly, variants flagged by the tool as likely sequencing artifacts — including FFPE artifacts, oxidative damage (8-oxoguanine), and other technical artifacts — were excluded if they showed low average tumor variant allele frequency (tVAF <15%). Variants from genomic positions with evidence of strand bias or likely to be of germline origin were also removed. After all filtering steps, 12,531 high-confidence somatic variants remained across the 1,086 samples.

Classification of hotspot mutations

Recurrent variants across tumor samples were identified, being defined as identical variants (same chromosome, position, reference, and alternate alleles) observed in more than one sample. Recurrent variants (hotspots) were retained if they were supported by somatic databases (e.g., COSMIC[36]) or located in known cancer driver genes. There were 115 unique hotspots (88 annotated to known driver genes) identified across the 1,086 samples.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Copy number data processing and QC

Somatic chromosome-level copy number alterations (CNAs) were estimated using read depth through a modified version of the tumor-only NGScopy assay (Bioconductor package NGScopy (version 1.17.1)[31]. Discrete copy number calls at gene-level were assigned to individual samples using the median absolute deviation (MAD), based on segmented copy number values across the RefGene dataset (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/, accessed 2020-08-17 19:54), containing 28,237 canonical transcripts[37]. Copy number calls (copy number ratio, CNR) were categorized based on their deviation from the MAD threshold, with negative scores indicating values 1 to 3 MAD below the median and positive scores indicating values 1 to 3 MAD above the median. For our analyses, we considered homozygous deletions (CNR of -2 to -3), amplifications (CNR of +2 to +3), and neutral (CNR of -1 to +1) copy number events at a gene-level.

QC assessment at sample level

The targeted sequencing was conducted on 1,235 OC and OPC tumor samples. Of these, 68 (5.5%) were of low quality, containing those with read depth of less than 30X, an average base quality score below 20, excessive duplicate reads, or inconsistent coverage patterns. Seventy-six (6.2%) samples were removed due to genotype mismatches between UNMASC germline calls and previously determined germline SNP-array data[38]. Additionally, a principal component analysis was conducted to identify samples enriched for sequencing artifacts that were identified by UNMASC and PICARD (version 3.3.0; https://github.com/broadinstitute/picard) (Figure S1); five (0.4%) outlying samples were removed based on these criteria. After applying these QC criteria, 1,086 tumor samples remained with mutation data, of which 808 also had corresponding high-quality CNA data. The patient group with both mutation and copy number data available (N=808) was comparable to the total study population (N=1,086) (Table S1).

Classification of frequently altered genes

Genes were classified as putative drivers if they exhibited positive selection, quantified by the maximum-likelihood nonsynonymous-to-synonymous ratio (dN/dS) using dNdScv (R v0.1.021)[39]. Samples with mutation counts exceeding three standard deviations from the median were not included to select genes under positive selection (N=62, 5.7%), and the dN/dS approach restricts the analyses to the somatic mutations present in the COSMIC database. Of

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

the 1,109 sequenced genes, 625 showed evidence of positive selection (dN/dS > 20, false discovery rate/FDR < 0.05 for non-synonymous substitutions) based on mutation data. Next, we combined mutations (either indels or single-nucleotide variants (SNVs)) with gene-level CNA events to classify 104 genes that were altered in at least 3% of all tumors. Finally, if multiple genes fell within the same CNA region, prioritization of the representative gene was based on (i) if a gene had been previously reported as a candidate or a driver of head and neck carcinogenesis[9,11] (e.g., *PIK3CA* [3q], *MYC* [8q], *CCND1* [11q], and *RB1* [13q]) or, (ii) the highest CNA frequency (Table S2). From this approach, 35 frequently altered genes were available for analysis. A schematic representation of the gene selection workflow is provided in Figure S2.

TCGA replication set

For the TCGA datasets, participant enrolment, collection of specimens, clinical and genomic data processing and distributions are subject to 45-CFR-46 (the "Common Rule") governing protection of human research subjects. Under the revised TCGA consent policy, re-consenting of still-living participants is no longer a programme-imposed requirement. The Project Team described the best practices for informed consent for participating in TCGA in this memo (http://cancergenome.nih.gov/abouttcga/policies/informedconsent).

Post-processed TCGA data was downloaded from cBioPortal (https://www.cbioportal.org/; accessed on April 8, 2025), restricted to OC (ICD10 code: C02.1, C02.9, C03.0, C03.1, C03.9, C04.0, C04.9, C05.0, C05.9, C06.0, C06.9) and OPC (C01, C09.9, C10.9) tumors. *CCND1* copy-number status (GISTIC2-categorized calls: AMP, NEUTRO, and DEL) and mRNA expression (RNA Seq V2 RSEM platform; z-scores relative to diploid samples) were retrieved. *TP53* mutation data were obtained for OC and OPC, and separately for TCGA lung adenocarcinoma (LUAD) tumors to provide a comparison group of a cancer type with an etiology related to both smoking and never-smoking. Harmonized demographic and clinical data, including HPV status (for OPC), tobacco and alcohol use, were also downloaded.

Statistical analyses

Differences in demographic and clinical variables and exposure histories between patient groups classified by tumor subsite and HPV status were assessed using chi-square or Fisher's exact tests. Age at diagnosis was categorized using 60 years as cutoff (mean ± sd: 59.8 ± 11.5 in total study population). For this study, somatic alterations included somatic mutations and CNA events. Alterations were tested in patients with both data types available (N=808), while

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

mutations were tested in all patients with mutation data available (N=1,086) and CNA events in all patients with CNA data available (N=808). Frequently altered genes and mutational hotspots were compared across groups classified by tumor subsite and HPV status (e.g., HPV(+)OPC, *vs*. HPV(-)OPC, and HPV(-)OPC *vs*. OC), and then further stratified by tobacco and alcohol use (ever/never) using Chi-square or Fisher's exact tests. For APOBEC-related mutations in *PIK3CA*, a mutation matrix containing *PIK3CA* mutations was generated using SigProfilerMatrixGenerator[40] (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) by applying the 'exome' option, which downsamples mutational matrices to the exome regions of the genome. The proportion of APOBEC-related mutations (T[C>T]A, T[C>T]T, T[C>G]A, and T[C>G]T)[41] in relation to all C>T and C>G mutations was calculated by tumor sample. Fisher's exact tests were used to explore differences in gene alteration frequencies between anatomical sites. For this, sites within the oral cavity were each compared to the tongue as the reference group; sites within the oropharynx were each compared to the tonsils as the reference group. Wilcoxon-Mann-Whitney tests were used to assess differences in non-synonymous mutation burden of cancer genes between tumor groups. Pairwise gene co-occurrence analysis was performed using the DISCOVER R package[42] (version 0.9.4), which accounts for sample-specific mutation rates and corrects for multiple testing. To assess the robustness of major findings and account for potential confounding, multivariable analyses were performed using Firth's penalized logistic regression to address data sparsity and prevent bias from monotone separation. Models evaluated included: (1) Model 1, univariate (unadjusted); (2) Model 2, adjusted for age, sex, and study (HN5000, ARCAGE, Toronto, CHANCE, Pittsburgh, INTERCHANGE); and (3) Model 3, adjusted for Model 2 variables and stage at diagnosis (Stages I–IV). The potential impact of the year of diagnosis on gene alteration frequencies was assessed as a continuous variable using Firth's penalized logistic regression, stratified by cancer subsite.

To account for multiple hypothesis testing, we applied the Benjamini-Hochberg procedure (false discovery rate/FDR), considering results statistically significant at FDR<5%. All statistical analyses were performed using R software (version 4.2.1).

Results

Study population

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Somatic mutations were characterized in the tumors from 1,086 OC and OPC patients. For 808 (74.4%), somatic CNA alterations were also inferred. Characteristics of the study population, stratified by cancer subsite and HPV status are shown in Table S3. The total study population includes 480 OPC patients [287 (59.8%) HPV(+)OPC, 174 (36.3%) HPV(-)OPC, and 19 (4.0%) with unknown HPV status] and 606 OC patients. Mean age at diagnosis was 59.8 (±11.5) years and the majority of the patients were male, ever smokers and ever drinkers (Table S1, Table S3). Compared to OC and HPV(-)OPC patients, HPV(+)OPC patients were more often younger at diagnosis, male and light-to-moderate smokers and drinkers, and less often both smoker and drinker (Table S3).

Analyses by subsite and HPV status reveal distinct somatic alteration patterns

Across 1,086 OC and OPC tumors, the 1,109 sequenced genes exhibited a mean non-synonymous mutation burden of 4.8 (SD: 11.0), which differed by HPV status ($P=2\times10^{-5}$) (Figure S3A). The highest mutation burden was observed in OC tumors (mean (SD): 5.4 (13.1)), followed by HPV(+)OPC (4.2 (8.7)) and HPV(-)OPC (3.5 (3.9); $P=9.4\times10^{-5}$).

We identified 35 frequently altered genes (Table S2), including 19 previously described as likely cancer drivers[9,12]. The distribution of alterations across these 19 driver genes, stratified by tumor subsite and HPV status, is shown in Figure 1. Among the known drivers, *TP53* (45.9%), *PIK3CA* (29.3%), and *CDKN2A* (19.1%) were the most frequently altered across all subsites. *TP53* and *CDKN2A* alterations were predominantly mutations and were especially common in HPV(-) tumors. *CCND1* (14.0%) and *EGFR* (6.7%) alterations, also more frequent in HPV(-) tumors, were primarily CNAs. In HPV(+)OPC, *PIK3CA* was the most frequently altered gene (44.9%), involving both mutations and CNAs, followed by *RB1* (14.1%; mostly deletions) and *FGFR3* (13.7%; mostly mutations) alterations. Figure S3 illustrates mutation hotspots in *TP53*, *PIK3CA*, *CDKN2A*, and *FGFR3* with differential frequencies between HPV(+) OPC and HPV(-) tumors; further details are provided in Table S4. Several frequently altered genes identified in our cohort were enriched in OC, including *RANBP17*, *SNX29*, *HECW1*, *ASXL1* (predominantly CNAs), and *NOTCH2* (predominantly mutations) (Table S2). Sensitivity analyses revealed no significant associations between the year of diagnosis and prevalence of the 35 most frequently altered genes across all subsites (all FDR > 0.05; Table S5).

PIK3CA alterations across HPV(+)OPC, HPV(-)OPC, and OC tumors

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

PIK3CA was the second most frequently altered gene overall and the most frequent in HPV(+)OPC (Figure 1, Table S2). While present in 18.3% of OC tumors, *PIK3CA* alterations were significantly more common in HPV(+)OPC (44.9%) and HPV(-)OPC (37.1%), with no significant difference in alteration frequency between HPV(+)OPC and HPV(-)OPC (Figure 2A, 2D, Table S2). Interestingly, the nature of these alterations varied: in HPV(-)OPC, *PIK3CA* changes were primarily amplifications, whereas in HPV(+)OPC they were predominantly somatic mutations, often occurring alongside amplifications at the *PIK3CA* locus (Figure 1, Figure 2).

Consistent with the viral etiology, *PIK3CA* mutations in HPV(+)OPC showed APOBEC-associated mutational profiles[41] and were enriched at the E542K and E545K hotspots (Figures 2B and 2C, Table S4). Notably, *FGFR3* mutations tended to co-occur with *PIK3CA* somatic mutations in HPV(+)OPC cases ($P=6.0\times10^{-7}$; Table S6). The *FGFR3* Ser249Cys gain-of-function hotspot mutation, also characterized by an APOBEC-associated mutational profile[4,9,43], was significantly enriched in HPV(+)OPC (5.2%, P=0.011) compared to OC (0.2%) and HPV(-)OPC (2.3%) (Table S4). This mutation hotspot frequently co-occurred with *PIK3CA* gain-of-function hotspots in codons 542 and 545 (Table S7). Together, these findings indicate that while 3q26/28 amplification is a hallmark of OPC, APOBEC-driven *PIK3CA* mutations, including the E542K and E545K hotspots, which often co-occur with *FGFR3* mutations in a subset of cases, are more specifically associated with HPV(+)OPC. Although less common in OC tumors, *PIK3CA* alterations are an important oncogenic driver in both OPC and OC tumors.

TP53 mutation frequencies across tumor subgroups and associations with smoking and alcohol consumption

TP53 was the most frequently altered gene in HPV(-) tumors, predominantly through somatic mutations (Figure 1, Table S2). As *TP53* mutations are hallmarks of tobacco-related carcinogenesis[9,44–47], we examined their relationship with smoking status across subsites. As expected, *TP53* mutations were less prevalent in HPV(+)OPC (Table S2), with no frequency difference between never (8.4%) and ever (6.0%) smokers (P=0.443; Table S8). In HPV(-)OPC, however, *TP53* mutations were significantly more common in ever smokers (64.2%) than in never (19.0%) smokers ($P=1.1\times10^{-04}$; Figure 3, Table 1, Table S8). This pattern is comparable to lung adenocarcinoma, where *TP53* mutation rates are 54.6% in ever and 38.4% in never smokers (P=0.007; Figure 3). Among HPV(-)OPC ever smokers, mutation frequency

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

was further stratified by smoking intensity, showing enrichment in heavy smokers (67.6%; ≥20 pack-years) compared to light-to-moderate smokers (46.7%; <20 pack-years) (P=0.032; Figure 3; Table S8).

In OC tumors, despite strong epidemiological evidence linking tobacco use to disease risk[1,9,12,13], *TP53* mutations were highly prevalent in both ever and never smokers (66.8% *vs.* 63.5%; P=0.444, Figure 3; Table S8). Interestingly, in OC never smokers, *CDKN2A* mutations significantly co-occurred with *TP53* mutations ($P=1.8\times10^{-04}$; Table S6). Because tobacco typically induces C>A transversions in *TP53*[9,12], we compared these proportions by smoking status. In both OC (11.2% in never *vs.* 11.4% ever smokers, P=1.00) and HPV(+)OPC (8.4% in never *vs.* 6.0% in ever smokers, P=0.846), proportions were similar (Table S9). By contrast, smoking-related C>A transversions in HPV(-)OPC were present in 14.9% of ever smokers but absent in never smokers (0%), which, albeit on small numbers, mirrored a pattern observed in lung adenocarcinoma (Figure 3, Table S9).

We further explored the relationship between combined tobacco and alcohol exposures and *TP53* alterations (Table S10). In HPV(-)OPC, *TP53* mutation frequency was highest among combined users (FDR=0.017), though small numbers in other categories limited further comparisons. In OC, where groups were larger and more balanced, *TP53* alterations remained uniformly high across exposure categories, with no evidence of cumulative effects (FDR=0.810).

To assess whether combined tobacco and alcohol exposures were associated with specific *TP53* mutation types (truncating *vs.* missense mutations), we evaluated each mutation type in HPV(-) tumors, combining both OC and HPV(-)OPC to increase power due to small sample sizes in some strata (Figure S4). Missense mutations were significantly more common than truncating mutations among ever smokers and ever drinkers (41.1% *vs.* 32.4%), ever smokers and never drinkers (39.0% *vs.* 25.0%), and never smokers and never drinkers (40.9% *vs.* 26.4%) (all $P < 0.05$). This pattern was not observed among never smokers who were ever drinkers, where the frequencies for missense and truncating mutations were similar. No significant differences in other mutation type frequencies were observed between exposure groups (Figure S4). Overall, these findings highlight that while *TP53* somatic mutations are common across HPV(-) head and neck tumors, its relationship with tobacco exposure, and the associated mutational processes, varies substantially by subsite.

Relationship between smoking and alcohol consumption and other frequently mutated genes

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

In addition to *TP53*, we also examined how somatic alterations in the other 34 frequently altered genes varied by tobacco and alcohol exposure (Table 1, S8 and S10). In HPV(-) tumors, smoking was linked to higher prevalence of *KMT2D* (12.8% *vs*. 5.3%) and *NSD1* (7.2% *vs*. 1.6%) alterations. Alcohol consumption showed similar relationships for *KMT2D* (13.2% vs. 5.2%) and *NSD1* (7.1% *vs*. 2.4%), whereas non-drinkers exhibited higher frequencies of *FAT1* (19.9% *vs*. 10.3%), *CASP8* (11.4% vs. 4.3%), *SNX29* (10.9% *vs*. 4.8%), *HECW1* (11.8% vs. 5.5%), and *HRAS* (10.0% vs. 4.3%) (all FDR < 0.05). Subsite-specific enrichments included *KMT2D* alterations in drinkers with OC (15.0% *vs*. 4.4%), and *FAT1* (20.3% *vs*. 10.5%), *CASP8* (12.1% *vs*. 4.8%), and *SNX29* (12.6% *vs*. 5.7%) alterations in non-drinkers with OC. Among HPV(-) tumors, further stratification of ever smokers by pack-years showed that heavy smokers had higher *CCND1* alteration frequencies (24.9% *vs*. 9.5%), while *HECW1* alterations were more frequent in light-to-moderate smokers, particularly in OC (15.3% *vs*. 3.6%).

With regard to differences by age and sex, we observed a few modest but significant differences in the alteration frequencies of the 35 genes across the full cohort (N=1,086; Table S8): 1) *CASP8* mutations were more frequent in older patients (≥ 60 years: 7.8% *vs*. < 60 years: 3.2%, FDR=0.030), and 2) *TP53* mutations were slightly more common in tumors from female patients than male patients (5.7% *vs*. 4.6%, FDR=0.039). However, when stratified by subsite, none of the differences remained significant (all FDR > 0.05; Table S8), suggesting that associations observed across the full cohort were driven by subsite distribution. Similarly, no significant differences in gene alteration frequencies were observed for stage at diagnosis within the subsites (all FDR > 0.05; Table S8).

Frequency of CCDN1 amplifications and relationship with smoking and alcohol consumption

CCND1 amplifications were primarily somatic amplifications at the 11q13 locus, occurring in 25.9% of HPV(-)OPC and 15.2% of OC, but were rare in HPV(+)OPC (4.0%) (Figure 1, Table 1). In HPV(-) tumors, the prevalence of *CCND1* amplifications was highest among smokers, particularly among those reporting combined tobacco and alcohol use compared to those with single or no exposure (FDR=0.017; Figure 4A, Table S10), suggesting a potential synergistic effect. Statistical testing for interaction between smoking and alcohol use by multivariable logistic regression using Firth's penalized likelihood and adjusting for age and sex showed evidence of multiplicative effects ($P_{interaction}$=0.003). To further investigate this observation, we analyzed an independent TCGA cohort of 290 HPV(-)OPC and OC tumors. Consistent with

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

our initial findings, *CCND1* amplifications were most frequent in individuals reporting both smoking and alcohol use (35.8%) compared to those with single or no exposures ($P=2.0\times10^{-04}$; Figure 4B). However, the interaction term did not reach significance ($P_{interaction}=0.790$). Integration with matched gene expression data from the TCGA cohort confirmed that this exposure-driven pattern translated to the transcriptomic level: *CCND1* expression was highest in tumors from dual users ($P=4\times10^{-4}$; $P_{interaction}=0.990$; Figure 4C), with elevated expression also observed in individuals reporting smoking or drinking alone. Together, these results suggest an exposure-driven relationship between smoking, drinking, and amplification-mediated *CCND1* expression in HPV(-) tumors, although the interaction results were not consistent across the two cohorts. The latter is likely due to differences in cohort size and / or population characteristics.

To ensure the robustness of the findings related to *PIK3CA*, *TP53*, and *CCND1*, among other genes, and to account for potential confounding, we evaluated the associations between gene alterations and primary variables of interest (cancer subsite, HPV status, smoking, and alcohol consumption) while adjusting for age, sex, study, and stage at diagnosis. The associations remained significant and largely unchanged after adjustment (Table S11), indicating that the observed genomic profiles are independent of potential demographic and clinical confounders.

Somatic alteration profiles by anatomic site

We next evaluated the distribution of somatic alterations in the 35 frequently altered genes across specific anatomic sites within OC and OPC (Figure 5, Table S12). Among OC tumors, smoking and alcohol use frequencies were highest in the floor of mouth (94% and 81%, respectively), while HPV(-)OPC tumors showed the highest exposure rates in the oropharynx (98% and 88%, respectively). These exposure patterns mirrored the high frequencies of *TP53*, *CDKN2A*, and *CCND1* alterations observed at these sites. Within OC, gingiva tumors were significantly enriched for *FGFR3* (14.3% *vs.* 3.2% in tongue; FDR=0.042) and *CASP8* alterations (17.9% *vs.* 2.9% in tongue; FDR=0.004) (Figure 5, Table S12).

Characterization of oral cavity tumors from never-smoker, never-drinker patients

Never smoker-never drinkers (NSND) accounted for 12.8% (N=139) of the total cohort (N=1,086). Within OC, we identified 100 NSND and 353 ever smoker-ever drinkers (ESED) (Table S13). Compared to ESED, NSND OC cases were more likely to be female (62% *vs.* 22%; $P=1.1\times10^{-13}$) and reported higher education levels ($P=6\times10^{-06}$). Geographically, the NSND profile was most prevalent in the Toronto study cohort[22] (49% *vs.* 2–18% elsewhere;

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

P=1.7×10^{-12}), potentially reflecting local differences and / or temporal shifts in tobacco and alcohol policies. No significant differences were observed for age or stage at diagnosis (all P > 0.05).

OC tumors from NSND exhibited a significantly higher prevalence of *FAT1* (21% *vs.* 10.5%; FDR=0.048) and *CASP8* mutations (14% *vs.* 4.2%; FDR=0.016) compared to those from ESED (Table S14, Figure 6). In contrast, *KMT2D* (16% *vs.* 3%; FDR=0.004) and *NSD1* (6.5% *vs.* 0%) alterations were enriched in ESED tumors compared to NSND. *TP53* and *CDKN2A* mutation frequencies did not differ significantly between the NSND and ESED tumor groups (Table S14, Figure 6; all FDR > 0.05).

Discussion

This study used one of the largest exposure-annotated cohorts of OC and OPC cases to date to investigate somatic alteration patterns, how these patterns relate to HPV status, and their relationship with tobacco and alcohol use.

Our findings underscore *TP53* as the main somatic driver in HPV(-) tumors. In HPV(-)OPC, *TP53* alterations were consistent with smoking-associated patterns observed in other tumor types [9,44–47], with higher prevalence among smokers, a gradient with increasing tobacco exposure and a tendency for C>A transversions among smokers. In contrast, in OC, the relationship was attenuated, *TP53* mutations were frequent in both ever and never smokers and there was no enrichment of C>A transversions among smokers. Consistent with this, a recent description of mutational signatures in OC tumors suggests a relative absence of the canonical tobacco SBS4 signature and, among smokers, a predominance of T>A transversions[12]. Taken together, these observations suggest that smoking may elevate OC risk via non-classical mutational processes, and that *TP53* inactivation in OC may arise through mechanisms distinct from canonical tobacco mutagenesis. Among patients without a history of alcohol use, OC tumors showed enrichment of *CASP8* mutations. These features are similar to previously described copy-neutral OC tumors[9,12,13,16], a subgroup associated with a more favorable prognosis[11]. How other etiological factors, such as poor oral hygiene, contribute to *TP53* alterations in OC may be worth exploring in independent studies.

Amplification of 11q13, a region containing *CCND1*, was a prominent feature of HPV(-) tumors in our cohort. *CCND1* encodes cyclin D1, a key regulator of the G1–S cell-cycle

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

transition, via activation of CDK4 and CDK6, whose overexpression drives uncontrolled proliferation[48]. *CCND1* amplification (11q13) is a common feature of squamous cell carcinomas[9,44,45,49,50] , including in tumors related to tobacco exposure[44,45]. Observational studies have reported a multiplicative effect of tobacco and alcohol on HNSCC risk[1–3]. Consistent with this synergy between tobacco and alcohol, the frequency of *CCND1* amplification in HPV(-) tumors was highest among patients with combined exposure. This association persisted at the transcript level with *CCND1* mRNA expression significantly elevated in dual users.

Mechanistically, acetaldehyde, a carcinogenic metabolite produced during ethanol metabolism and present in tobacco smoke, may upregulate *CCND1* in OC and HPV(-)OPC tumors by inducing genomic instability that promotes copy-number amplification. It may also activate ERK/AP-1 and NF-κB signaling, thereby enhancing cyclin D1 transcription via its promoter response elements[51–55]. Clinically, *CCND1* amplification and overexpression have been linked to tumor progression and poor prognosis[56,57], particularly in OC tumors. These findings suggest that HPV(-) tumor patients with both exposures could benefit from therapies targeting the CDK4/6 pathway, agents currently under investigation for HPV(-) head and neck cancer, and reinforce the importance of smoking cessation, given the established survival advantage among quitters[58]. Together, these results underscore the biological and clinical importance of tobacco and alcohol as co-drivers of aggressive HPV(-) HNSCC.

PIK3CA is a central oncogenic driver in OPC tumors, but its role varies depending on HPV status and the underlying mutational processes. Our data confirm that *PIK3CA* is frequently altered in HPV(+)OPC, as previously described in both exome and targeted sequencing studies[4,9,43]. However, we demonstrate that the nature of *PIK3CA* alterations is context dependent. Amplification of the 3q26/28 region, encompassing *PIK3CA* but also other relevant genes that play key roles in maintaining the self-renewal capacity of cancer stem cells in head and neck tumors (*SOX2*[59]) and the squamous cell identity (*TP63*[60]), was common in both HPV(+) and HPV(-) OPCs, consistent with TCGA findings[9]. This region is known to be a major target of genomic instability in squamous cancers, underscoring a shared, subsite-specific mechanism of OPC tumorigenesis beyond viral status. In contrast, gain of function point mutations in *PIK3CA*, especially the canonical E545K and E542K hotspots, were significantly more common in HPV(+)OPC and showed clear signatures of APOBEC-mediated mutagenesis[4,8,9,41,43]. This is consistent with the well-established role of APOBEC enzymes in shaping the mutational landscape of HPV-associated tumors[4,9,11]. Interestingly, we also noted a

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

subset of HPV(+)OPC with APOBEC-related gain-of-function mutations in both *PIK3CA* and *FGFR3*, supporting prior evidence of coordinated APOBEC activity driving key oncogenic events in this context[8,41]. Taken together, these findings highlight two distinct but overlapping mechanisms for PI3K pathway activation in OPC: (a) focal 3q amplification (in both HPV(+) and HPV(-) OPC) and, (b) APOBEC-driven hotspot mutation (predominantly in HPV(+)OPC). Clinically, this distinction may have implications for the use of PI3K-targeted therapies and emphasizes the need to account for both mutational and copy number alterations when considering molecular stratification or therapeutic targeting of the PI3K pathway in HNSCC.

Beyond broad subsite comparisons, our cohort size enabled the first exploration of driver genes specific to anatomical sites within OC and OPC. Notably, we found that among OC cases, *FGFR3* and *CASP8* alterations were significantly enriched in gingival tumors. While the biological mechanisms underlying these site-specific enrichments remain unclear, possible explanations include unique inflammatory exposures, microenvironmental differences, or epithelial cell composition in gingiva compared to other oral cavity subsites. These findings highlight the importance of considering anatomical site in future studies of HNSCC somatic alterations.

The major patterns of somatic driver alterations in our cohort, including alterations in *TP53*, *CDKN2A*, *CCND1* and *PIK3CA* closely mirror observations from previous large-scale studies, reinforcing their central roles across HNSCC subsites and exposure groups. We also observed suggestive trends with several less-characterized genes, such as *HECW1* and *SNX29*, as frequently altered in our dataset, and showing exposure- or subsite-specific associations not previously reported in HNSCC.

Several limitations should be noted. The use of tumor-only sequencing without matched normal controls may have limited the ability to distinguish rare germline from true somatic variants. However, we used a validated pipeline specifically designed for this context, incorporating public somatic and germline databases to minimize misclassification. In addition, while targeted sequencing does not capture genome-wide mutation or copy number variation, it provides high-depth, robust profiling of key cancer driver genes. Finally, the cross-sectional design of this study precludes a temporal evaluation of early carcinogenesis steps, as evaluating tumors exclusively at diagnosis prevents us from determining whether specific driver alterations arose from age-related accumulation or subsequent exposure-driven clonal expansion[61,62]. To mitigate this issue, our analysis focused on established cancer driver genes

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

using a targeted 1,109-gene panel, rather than broader passenger mutations. Nevertheless, prospective longitudinal cohorts are required to confirm whether monitoring or intercepting these specific somatic alterations can effectively prevent or detect head and neck cancers early. Major strengths include our large, multinational, exposure-annotated cohort, which complements and extends the current literature by providing molecular data from a more diverse patient population.

In summary, this study demonstrates that the molecular architecture of OC and OPC is shaped by tumor subsite, HPV status, and patterns of tobacco and alcohol exposures. The effects of these risk factors on pivotal genes such as *TP53* and *CCND1* are highly subsite-dependent, with evidence of synergism between tobacco and alcohol for the latter. Collectively, these findings highlight the molecular heterogeneity of HNSCC, support a shift toward more tailored molecular stratification, and underscore the need for prevention strategies that account for both environmental exposures and tumor biology.

Data availability

The data underlying this article, including processed tumor somatic variant calling of 1,086 tumour samples and associated metadata, will be available at dbGAP under accession number phs002349.v1.p1.

Disclosure and competing interests statement

Tim Waterboer serves on advisory boards for MSD (Merck) Sharp & Dohme. Scott Bratman is inventor on patents related to cell-free DNA mutation and methylation analysis technologies that are unrelated to this work and have been licensed to Roche and Adela, respectively, and is a co-founder of and has ownership in Adela. The authors declare no competing interests. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policies, or views of the International Agency for Research on Cancer/World Health Organization. Portions of the manuscript text were edited for clarity and grammar using generative AI. All authors reviewed and approved the final wording and take full responsibility for the content.

Acknowledgement

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

The authors would like to thank all study participants. The funder had no role in the design of the study; the collection, analysis, or interpretation of the data; or the writing of the manuscript and decision to submit it for publication.

Funding

This research was supported by US National Institute of Dental and Craniofacial Research (NIDCR) grants R01DE025712 and R03DE030257. The Alcohol-Related Cancers and Genetic Susceptibility Study in Europe (ARCAGE) was funded by the European Commission's fifth frame-work program (QLK1-2001-00182), the Italian Association for Cancer Research, Compagnia di San Paolo/FIRMS, Region Piemonte and Padova University (CPDA057222). The Carolina Head and Neck Cancer Epidemiology (CHANCE) study was supported in part by the US National Cancer Institute (R01CA90731). The Head and Neck 5000 study was a component of independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10034). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Core funding was also provided through awards from Above and Beyond, University Hospitals Bristol and Weston Research Capability Funding and the NIHR Senior Investigator award to Professor Andy Ness. Interchange Brazil was funded by FAPES (328/2022). Human papilloma virus (HPV) serology was supported by a Cancer Research UK Programme Grant, the Integrative Cancer Epidemiology Programme (grant number: C18281/A19169). The University of Pittsburgh head and neck cancer case-control study was supported by US National Institutes of Health grants P50CA097190 and P30CA047904. The MSH-PMH study was supported by the Canadian Cancer Society Research Institute and the Princess Margaret Head & Neck Translational Research Program, with philanthropic funds from Joe's Team and the Wharton, Elia, Riley, and Tozer families. G.L. is funded by the Alan B. Brown Chair in Molecular Genomics and the Lusi Wong Foundation Fund. Funding was also provided by the International Agency for Research on Cancer and the European Union's Horizon 2020 research and innovation program HEADSpAcE under Grant Agreement No. 825771.

References

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

1. Gormley M, Creaney G, Schache A, Ingarfield K, Conway DI. Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *Br Dent J*. 2022;233(9):780-786. doi:10.1038/s41415-022-5166-x

2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263. doi:10.3322/caac.21834

3. Hashim D, Genden E, Posner M, Hashibe M, Boffetta P. Head and neck cancer prevention: from primary prevention to impact of clinicians on reducing burden. *Annals of Oncology*. 2019;30(5):744-756. doi:10.1093/annonc/mdz084

4. Lechner M, Liu J, Masterson L, Fenton TR. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nat Rev Clin Oncol*. 2022;19(5):306-327. doi:10.1038/s41571-022-00603-7

5. Ang KK, Harris J, Wheeler R, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *New England Journal of Medicine*. 2010;363(1):24-35. doi:10.1056/NEJMoa0912217

6. Rettig EM, Zaidi M, Faraji F, et al. Oropharyngeal cancer is no longer a disease of younger patients and the prognostic advantage of Human Papillomavirus is attenuated among older patients: Analysis of the National Cancer Database. *Oral Oncol*. 2018;83:147-153. doi:10.1016/j.oraloncology.2018.06.013

7. Hayes DN, Van Waes C, Seiwert TY. Genetic Landscape of Human Papillomavirus–Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors. *Journal of Clinical Oncology*. 2015;33(29):3227-3234. doi:10.1200/JCO.2015.62.1086

8. Cannataro VL, Gaffney SG, Sasaki T, et al. APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene*. 2019;38(18):3475-3487. doi:10.1038/s41388-018-0657-6

9. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576-582. doi:10.1038/nature14129

10. Chung CH, Guthrie VB, Masica DL, et al. Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing. *Annals of Oncology*. 2015;26(6):1216-1223. doi:10.1093/annonc/mdv109

11. Muijlwijk T, Nauta IH, van der Lee A, et al. Hallmarks of a genomically distinct subclass of head and neck cancer. *Nat Commun*. 2024;15(1):9060. doi:10.1038/s41467-024-53390-3

12. Torrens L, Moody S, de Carvalho AC, et al. The complexity of tobacco smoke-induced mutagenesis in head and neck cancer. *Nat Genet*. 2025;57(4):884-896. doi:10.1038/s41588-025-02134-0

13. Ghasemi F, Prokopec SD, MacNeil D, et al. Mutational analysis of head and neck squamous cell carcinoma stratified by smoking status. *JCI Insight*. 2019;4(1). doi:10.1172/jci.insight.123443

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

14. Papenberg BW, Ingles J, Gao S, et al. Copy number alterations identify a smoking-associated expression signature predictive of poor outcome in head and neck squamous cell carcinoma. *Cancer Genet*. 2021;256-257:136-148. doi:10.1016/j.cancergen.2021.05.011

15. Li Z, Qiu X, He Q, Fu X, Ji F, Tian X. CCND1-associated ceRNA network reveal the critical pathway of TPRG1-AS1-hsa-miR-363-3p-MYO1B as a prognostic marker for head and neck squamous cell carcinoma. *Sci Rep*. 2023;13(1):11831. doi:10.1038/s41598-023-38847-7

16. Koo K, Mouradov D, Angel CM, et al. Genomic Signature of Oral Squamous Cell Carcinomas from Non-Smoking Non-Drinking Patients. *Cancers (Basel)*. 2021;13(5):1029. doi:10.3390/cancers13051029

17. Lechner M, Frampton GM, Fenton T, et al. Targeted next-generation sequencing of head and neck squamous cell carcinoma identifies novel genetic alterations in HPV+ and HPV- tumors. *Genome Med*. 2013;5(5):49. doi:10.1186/gm453

18. Gormley M, Adhikari A, Dudding T, et al. VOYAGER: an international consortium investigating the role of human papilloma virus and genetics in oral and oropharyngeal cancer risk and survival. Preprint posted online February 21, 2025. doi:10.1101/2025.02.17.25322399

19. Abrahão R, Perdomo S, Pinto LFR, et al. Predictors of Survival After Head and Neck Squamous Cell Carcinoma in South America: The InterCHANGE Study. *JCO Glob Oncol*. 2020;(6):486-499. doi:10.1200/GO.20.00014

20. Lagiou P, Georgila C, Minaki P, et al. Alcohol-related cancers and genetic susceptibility in Europe: the ARCAGE project: study samples and data collection. *European Journal of Cancer Prevention*. 2009;18(1):76-84. doi:10.1097/CEJ.0b013e32830c8dca

21. Ness AR, Waylen A, Hurley K, et al. Establishing a large prospective clinical cohort in people with head and neck cancer as a biomedical resource: head and neck 5000. *BMC Cancer*. 2014;14(1):973. doi:10.1186/1471-2407-14-973

22. Thomas S, Carroll JC, Brown MC, et al. Nicotine dependence as a risk factor for upper aerodigestive tract (UADT) cancers: A mediation analysis. *PLoS One*. 2020;15(8):e0237723. doi:10.1371/journal.pone.0237723

23. Troy JD, Grandis JR, Youk AO, Diergaarde B, Romkes M, Weissfeld JL. Childhood passive smoke exposure is associated with adult head and neck cancer. *Cancer Epidemiol*. 2013;37(4):417-423. doi:10.1016/j.canep.2013.03.011

24. Divaris K, Olshan AF, Smith J, et al. Oral health and risk for head and neck squamous cell carcinoma: the Carolina Head and Neck Cancer Study. *Cancer Causes & Control*. 2010;21(4):567-575. doi:10.1007/s10552-009-9486-9

25. Kusters JMA, Diergaarde B, Ness A, et al. Diagnostic accuracy of <scp>HPV16</scp> early antigen serology for <scp>HPV</scp> -driven oropharyngeal cancer is independent of age and sex. *Int J Cancer*. 2024;154(2):389-402. doi:10.1002/ijc.34710

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

26. Hibbert J, Halec G, Baaken D, Waterboer T, Brenner N. Sensitivity and Specificity of Human Papillomavirus (HPV) 16 Early Antigen Serology for HPV-Driven Oropharyngeal Cancer: A Systematic Literature Review and Meta-Analysis. *Cancers (Basel)*. 2021;13(12):3010. doi:10.3390/cancers13123010

27. Ito Fukunaga M, Wiener RS, Slatore CG. The 2021 US Preventive Services Task Force Recommendation on Lung Cancer Screening. *JAMA Oncol*. 2021;7(5):684. doi:10.1001/jamaoncol.2020.8376

28. Crowley R, Hilden D, Beachy M, et al. Excessive Alcohol Use and Alcohol Use Disorders: A Policy Brief of the American College of Physicians. *Ann Intern Med*. 2024;177(5):656-657. doi:10.7326/M23-3434

29. Little P, Jo H, Hoyle A, et al. UNMASC: tumor-only variant calling with unmatched normal controls. *NAR Cancer*. 2021;3(4). doi:10.1093/narcan/zcab040

30. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*. 2017;45(D1):D499-D506. doi:10.1093/nar/gkw879

31. Zhao X, Wang A, Walter V, et al. Combined Targeted DNA Sequencing in Non-Small Cell Lung Cancer (NSCLC) Using UNCseq and NGScopy, and RNA Sequencing Using UNCqeR for the Detection of Genetic Aberrations in NSCLC. *PLoS One*. 2015;10(6):e0129280. doi:10.1371/journal.pone.0129280

32. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

33. Tischler G, Leonard S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol Med*. 2014;9(1):13. doi:10.1186/1751-0473-9-13

34. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*. 2014;30(19):2813-2815. doi:10.1093/bioinformatics/btu376

35. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15(8):591-594. doi:10.1038/s41592-018-0051-x

36. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

37. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res*. 2021;49(D1):D884-D891. doi:10.1093/nar/gkaa942

38. Ebrahimi E, Sangphukieo A, Park HA, et al. Cross-ancestral GWAS identifies 29 variants across head and neck cancer subsites. *Nat Commun*. 2025;16(1):8787. doi:10.1038/s41467-025-63842-z

39. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-1041.e21. doi:10.1016/j.cell.2017.09.042

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

40. Khandekar A, Vangara R, Barnes M, et al. Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. *BMC Genomics*. 2023;24(1):469. doi:10.1186/s12864-023-09584-y

41. Petljak M, Dananberg A, Chu K, et al. Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature*. 2022;607(7920):799-807. doi:10.1038/s41586-022-04972-y

42. Canisius S, Martens JWM, Wessels LFA. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol*. 2016;17(1):261. doi:10.1186/s13059-016-1114-x

43. Seiwert TY, Zuo Z, Keck MK, et al. Integrative and Comparative Genomic Analysis of HPV-Positive and HPV-Negative Head and Neck Squamous Cell Carcinomas. *Clinical Cancer Research*. 2015;21(3):632-641. doi:10.1158/1078-0432.CCR-13-3310

44. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi:10.1038/nature11404

45. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550. doi:10.1038/nature13385

46. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 Mutation, and Lung Cancer. *Molecular Cancer Research*. 2014;12(1):3-13. doi:10.1158/1541-7786.MCR-13-0539

47. de Andrade KC, Lee EE, Tookmanian EM, et al. The TP53 Database: transition from the International Agency for Research on Cancer to the US National Cancer Institute. *Cell Death Differ*. 2022;29(5):1071-1073. doi:10.1038/s41418-022-00976-3

48. Fassl A, Geng Y, Sicinski P. CDK4 and CDK6 kinases: From basic science to cancer therapy. *Science (1979)*. 2022;375(6577). doi:10.1126/science.abc1495

49. Dotto GP, Rustgi AK. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. *Cancer Cell*. 2016;29(5):622-637. doi:10.1016/j.ccell.2016.04.004

50. Integrated genomic characterization of oesophageal carcinoma. *Nature*. 2017;541(7636):169-175. doi:10.1038/nature20805

51. Pawlonka J, Rak B, Ambroziak U. The regulation of cyclin D promoters - review. *Cancer Treat Res Commun*. 2021;27:100338. doi:10.1016/j.ctarc.2021.100338

52. Guttridge DC, Albanese C, Reuther JY, Pestell RG, Baldwin AS. NF-kappaB controls cell growth and differentiation through transcriptional regulation of cyclin D1. *Mol Cell Biol*. 1999;19(8):5785-5799. doi:10.1128/MCB.19.8.5785

53. Tetsu O, McCormick F. Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*. 1999;398(6726):422-426. doi:10.1038/18884

54. Román J, Giménez A, Lluis JM, et al. Enhanced DNA binding and activation of transcription factors NF-kappa B and AP-1 by acetaldehyde in HEPG2 cells. *J Biol Chem*. 2000;275(19):14684-14690. doi:10.1074/jbc.275.19.14684

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

55. Román J, Colell A, Blasco C, et al. Differential role of ethanol and acetaldehyde in the induction of oxidative stress in HEP G2 cells: effect on transcription factors AP-1 and NF-kappaB. *Hepatology*. 1999;30(6):1473-1480. doi:10.1002/hep.510300623

56. Ramos-García P, González-Moles MÁ, Ayén Á, González-Ruiz L, Gil-Montoya JA, Ruiz-Ávila I. Predictive value of *CCND1* /cyclin D1 alterations in the malignant transformation of potentially malignant head and neck disorders: Systematic review and meta-analysis. *Head Neck*. 2019;41(9):3395-3407. doi:10.1002/hed.25834

57. Kaminagakura E, Werneck da Cunha I, Soares FA, Nishimoto IN, Kowalski LP. *CCND1* amplification and protein overexpression in oral squamous cell carcinoma of young patients. *Head Neck*. 2011;33(10):1413-1419. doi:10.1002/hed.21618

58. Caini S, Del Riccio M, Vettori V, et al. Post-diagnosis smoking cessation and survival of patients with head and neck cancer: a systematic review and meta-analysis. *Br J Cancer*. 2022;127(11):1907-1915. doi:10.1038/s41416-022-01945-w

59. Lee SH, Oh SY, Do SI, et al. SOX2 regulates self-renewal and tumorigenicity of stem-like cells of head and neck squamous cell carcinoma. *Br J Cancer*. 2014;111(11):2122-2130. doi:10.1038/bjc.2014.528

60. Dotto GP, Rustgi AK. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. *Cancer Cell*. 2016;29(5):622-637. doi:10.1016/j.ccell.2016.04.004

61. Yoshida K. Somatic mutations and clonal evolution in normal tissues and cancer development. *Exp Mol Med*. 2026;58(4):961-969. doi:10.1038/s12276-025-01592-0

62. Hill W, Lim EL, Weeden CE, et al. Lung adenocarcinoma promotion by air pollutants. *Nature*. 2023;616(7955):159-167. doi:10.1038/s41586-023-05874-3

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure legends

Figure 1. Overall view of the alteration types in driver genes in oral cavity (OC) and oropharyngeal (OPC) tumors. Waterfall plot representing the types of alterations in driver genes ranked by the overall frequency in the study population. Samples were split by cancer subsites (HPV(-) OPC, N=143; OC, N=421; HPV(+) OPC, N=227) within (A) OC and HPV(-)OPC and (B) HPV(+)OPC. Somatic mutations (hotspots in black, missense in mustard, truncating in green, and silent alterations in grey) and homozygous copy number alterations (CNA; amplification/AMP in burgundy and deletions/DEL in blue colours) by gene (rows) and by sample (cols). On the right side, the genomic location (cytoband), overall frequencies within strata as barplots and cancer subsite-stratified frequencies are shown. Upper side annotations included age (≥60 years in blue *vs.* <60 years old in light yellow), sex (female in blue *vs.* male in light yellow), HPV status (positive in blue *vs.* negative in light yellow), and cancer subsite (OC in light yellow and OPC in blue). On the left side, gene names were flagged in bold font when their frequencies differ between HPV(-) OPC *vs.* OC in (A), or between HPV(+) OPC *vs.* HPV(-) tumors (HPV(-) OPC+OC) in (B) (Two-sided Fisher's exact test, after multiple testing correction (false discovery rate/FDR<0.05, n=35 tests). Of the list of 35 frequently altered genes, we show the 19 known or likely driver genes, previously reported by The Cancer Genome Atlas[9] or Mutographs'[12] papers in head and neck tumors. Further details on the frequencies of the 35 genes by cancer subsite are in Table S2. For representation purposes, we considered, whenever detected, homozygous amplifications for oncogenes and homozygous deletions for tumor suppressor genes.

Figure 2. Distribution of *PIK3CA* alterations by cancer subsite. Bar plots show the frequency of *PIK3CA* alterations across HPV-positive oropharyngeal cancer (HPV(+)OPC), HPV-negative OPC (HPV(-)OPC), and oral cavity (OC) tumors. (A) Combined *PIK3CA* amplifications and mutations (Yes: yellow; No: blue) in total N=791 tumors. (B) APOBEC mutations in *PIK3CA* (Yes: pink; No: blue; 'T[C>T]A', 'T[C>T]T', 'T[C>G]A', 'T[C>G]T') compared to all C>T and C>G mutation contexts in total N=1067 tumors. (C) Most frequent hotspot mutations (Glu545Lys and Glu542Lys) (Yes: pink; No: blue) in total N=1067 tumors. (D) *PIK3CA* amplifications (3q26/28) only where CNA were available (666 out of 791 CNA available) (Yes: purple; No: green). P-values were derived from two-sided Fisher's exact tests using OC as the reference group. Pairwise comparisons between HPV(+)OPC and HPV(-)OPC were also conducted. Significance is indicated as follows: *p<0.05, **p<0.01, ***p<0.001, NS (not significant, p>0.05).

Figure 3. Distribution of *TP53* alterations by smoking status and intensity across tumor types. Each barplot displays the proportion and absolute number of tumors with *TP53* mutations (MUT) or wild-type *TP53* (WT), stratified by smoking status (ever *vs.* never). Tumor types shown include (A) HPV-positive OPC (HPV(+)OPC; negative control), (B) HPV-negative oropharyngeal cancer (HPV(-)OPC), (C) TCGA lung adenocarcinoma (LUAD; positive control), and (E) oral cavity (OC). Smoking groups are divided into never-smoker *TP53*-MUT (dark yellow), never-smoker *TP53*-WT (light yellow), ever-smoker *TP53*-MUT (dark blue), and ever-smoker *TP53*-WT (light blue). Percentages and sample counts are shown for each category. P-values from Fisher's exact test compare mutation frequencies (MUT *vs.* WT) between ever and never smokers. (D) For HPV(-) OPC, ever smokers are further stratified by smoking intensity (<20 *vs.* ≥20 pack-years, PY). Mutation status is shown for light-to-moderate smokers (dark/light green) and heavy smokers (dark/light purple). Significance is

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

indicated as follows: *p<0.05, **p<0.01, ***p<0.001, NS (not significant, p>0.05). Abbreviations: HPV, human papillomavirus; MUT, mutant; WT, wild-type; PY, pack-years.

Figure 4. *CCND1* amplifications and expression by tobacco and alcohol exposure in HPV-negative tumors. Frequencies of *CCND1* amplifications and associated 95% confidence intervals are shown across combinations of smoking and alcohol use in oral cavity (OC) and HPV-negative oropharyngeal (OPC) tumors from **(A)** our study and **(B)** The Cancer Genome Atlas (TCGA) cohort. Pairwise P-values in panel A and B were calculated using two-sided Fisher's exact tests, with never smokers and never drinkers (NeverSmk;NeverDrk) as the reference group. **(C)** Boxplot of $\log_2$-transformed *CCND1* expression levels in OC and HPV-negative OPC tumors from TCGA. The plot displays the median, interquartile range (IQR: 25th–75th percentiles), whiskers (1.5× IQR), and outliers. P-values in panel C were derived from pairwise Wilcoxon tests using the NeverSmk;NeverDrk group as reference. $P_{interaction}$ from panels A-C were calculated using regression models (*CCND1* ~ Smk+Drk+Smk*Drk), where P-values are derived from the interaction term. Smoking and drinking subgroups are color-coded as follows: Ever smokers and ever drinkers (EverSmk;EverDrk) in orange; ever smokers and never drinkers (EverSmk;NeverDrk) in light blue; never smokers and ever drinkers (NeverSmk;EverDrk) in green; never smokers and never drinkers (NeverSmk;NeverDrk) in dark grey. Statistical significance is indicated as: *p<0.05, **p<0.01, ***p<0.001.

Figure 5. Distribution of the driver genes between oral cavity (OC) and oropharyngeal (OPC) anatomical sites. Heatmap representing frequencies of the 19 driver genes across anatomical sites of OC and OPC. OPC tumors were split by human papilloma virus (HPV) status (HPV(-)OPC and HPV(+)OPC). Fisher's exact tests evaluated differences in the alteration frequencies across anatomical sites within OC (buccal+retromolar: n=75; floor of mouth/FoM: n=100; gingiva: n=56, palate:n=38; tongue: n=309) and oropharyngeal/OPC (HPV(-)base of tongue/BoT: n=36; HPV(-) oropharynx: n=43; HPV(-)tonsils: n=64; HPV(+)BoT: n=78; HPV(+)oropharynx: n=26; HPV(+)tonsils: n=179. For the statistical tests, the reference groups were tongue for OC and tonsils for OPC. Gene names on the right side as rows and anatomical sites as columns. The intensity of red corresponds to frequency of gene alterations. * when false discovery rate/FDR<0.05 and ** when FDR<0.01 (n=35 tests). Bottom annotations represent the frequencies of ever smokers (Smk) and ever drinkers (Drk) by anatomical site. Of the list of 35 frequently altered genes, we represented the 19 known driver genes, previously reported by The Cancer Genome Atlas[9] or Mutographs'[12] papers in head and neck tumors. Further details on the frequencies of the 35 genes (Table S2, Table S12).

Figure 6. Genomic profile of NSND OC tumors. Comparison of somatic alteration frequencies (combined mutations and copy number alterations) between never smoker-never drinker (NSND; N=100) and ever smoker-ever drinker (ESED; N=353) oral cavity (OC) tumors. Genes are ordered along the y-axis by their alteration frequency within the ESED subgroup. Percentages indicate the prevalence of alterations for each gene within the respective strata. Statistical significance was assessed using two-sided Fisher's exact tests with Benjamini-Hochberg (FDR) correction across the 19 known driver genes shown. Brackets and annotated FDR values indicate significant differences between groups; these comparisons were restricted to genes with more than one alteration event per group to ensure biological relevance. An asterisk (*) denotes genes where zero alteration events were observed in one clinical group (e.g., *NSD1*).

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

892 **Table 1.** Relationship between HPV status, and smoking and drinking histories, and somatic alterations

Exposure	Gene	Strata	Freq (%) (95% CI)	FDR[b]	Subsite[c]
HPV Status	*TP53*	Positive (287)	7.3% (4.6%, 11%)	2.7×10^{-32}	OPC
		Negative (174)	58.6% (50.9%, 66%)		
	CCND1[d]	Positive (227)	4.0% (1.8%, 7.4%)	1.4×10^{-08}	
		Negative (143)	25.9% (18.9%, 33.9%)		
	CDKN2A	Positive (287)	5.6% (3.2%, 8.9%)	0.001	
		Negative (174)	17.2% (11.9%, 23.7%)		
	EGFR	Positive (287)	2.1% (0.8%, 4.5%)	0.032	
		Negative (174)	8% (4.5%, 13.1%)		
	PTEN	Positive (287)	9.8% (6.6%, 13.8%)	0.034	
		Negative (174)	2.9% (0.9%, 6.6%)		
Tobacco Smoking	*TP53*	Never (21)	19.0% (5.4%, 41.9%)	0.004	HPV(-)OPC
		Ever (148)	64.2% (55.9%, 71.9%)		
	KMT2D	Never (188)	5.3% (2.6%, 9.6%)	0.041	HPV(-) tumors
		Ever (585)	12.8% (10.2%, 15.8%)		
	NSD1	Never (188)	1.6% (0.3%, 4.6%)	0.041	
		Ever (585)	7.2% (5.2%, 9.6%)		
Alcohol Drinking	*KMT2D*	Never (182)	4.4% (1.9%, 8.5%)	0.003	OC
		Ever (420)	15.0% (11.7%, 18.8%)		
	FAT1	Never (182)	20.3% (14.7%, 26.9%)	0.030	
		Ever (420)	10.5% (7.7%, 13.8%)		
	CASP8	Never (182)	12.1% (7.7%, 17.7%)	0.030	
		Ever (420)	4.8% (2.9%, 7.3%)		
	SNX29	Never (182)	12.6% (8.2%, 18.4%)	0.043	
		Ever (420)	5.7% (3.7%, 8.4%)		
	CASP8	Never (211)	11.4% (7.4%, 16.5%)	0.012	HPV(-) tumors
		Ever (561)	4.3% (2.8%, 6.3%)		
	FAT1	Never (211)	19.9% (14.7%, 25.9%)	0.012	
		Ever (561)	10.3% (7.9%, 13.2%)		
	KMT2D	Never (211)	5.2% (2.6%, 9.1%)	0.014	
		Ever (561)	13.2% (10.5%, 16.3%)		
	HECW1	Never (211)	11.8% (7.8%, 17%)	0.025	
		Ever (561)	5.5% (3.8%, 7.8%)		
	SNX29	Never (211)	10.9% (7%, 15.9%)	0.025	
		Ever (561)	4.8% (3.2%, 6.9%)		
	HRAS	Never (211)	10.0% (6.3%, 14.8%)	0.025	
		Ever (561)	4.3% (2.8%, 6.3%)		
	NSD1	Never (211)	2.4% (0.8%, 5.4%)	0.042	
		Ever (561)	7.1% (5.1%, 9.6%)		
Pack-years	*HECW1*	Light-moderate (98)	15.3% (8.8%, 24.0%)	0.007	OC
		Heavy (279)	3.6% (1.7%, 6.5%)		
	HECW1	Light-moderate (128)	13.3% (7.9%, 20.4%)	0.017	HPV(-) tumors
		Heavy (381)	3.9% (2.2%, 6.4%)		

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

	CCND1[d]	Light-moderate (95)	9.5% (4.4%, 17.2%)	0.021	
		Heavy (273)	24.9% (19.9%, 30.5%)		

[a]Only relationships that reached significance threshold (FDR<5%) are represented in this table. All results reported in Table S8. No statistically significant enrichment observed for age (>60 vs. ≤60years), sex, drinking intensity, and stage at diagnosis at FDR (false discovery rate) <5%. No association between smoking and drinking status among HPV(+)OPC cases at FDR<5%. [b]FDR <5% P-values are reported. [c]HPV(-) tumors: OC and HPV(-)OPC combined. [d]Analyses with *CCND1* were here restricted to samples with complete CNA and mutation data since most alterations in *CCND1* were amplifications.

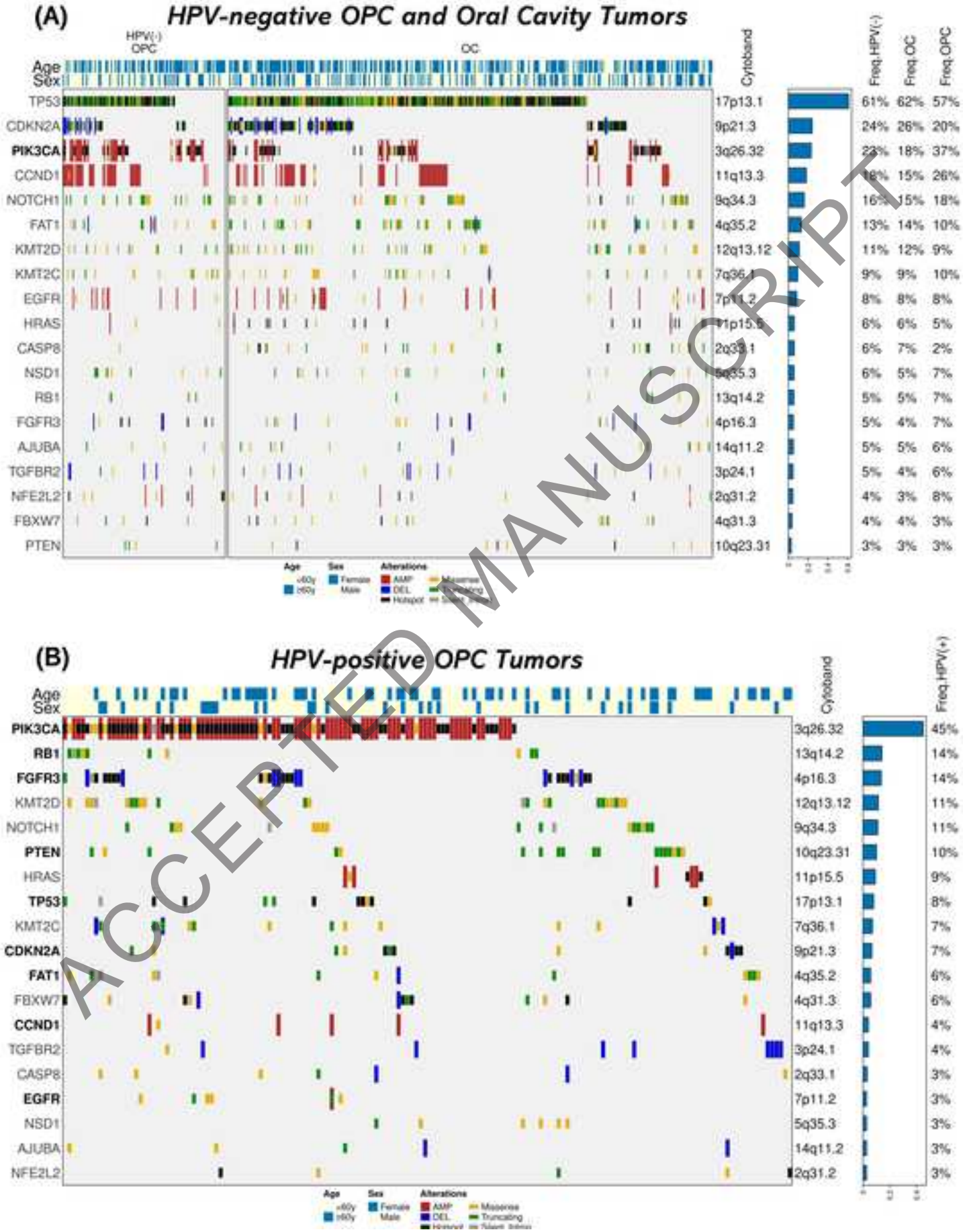
ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 1. Overall view of the alteration types in driver genes in oral cavity (OC) and oropharyngeal (OPC) tumors.

Click here to access/download;Figure - Color;Figure1.tif

Figure 1

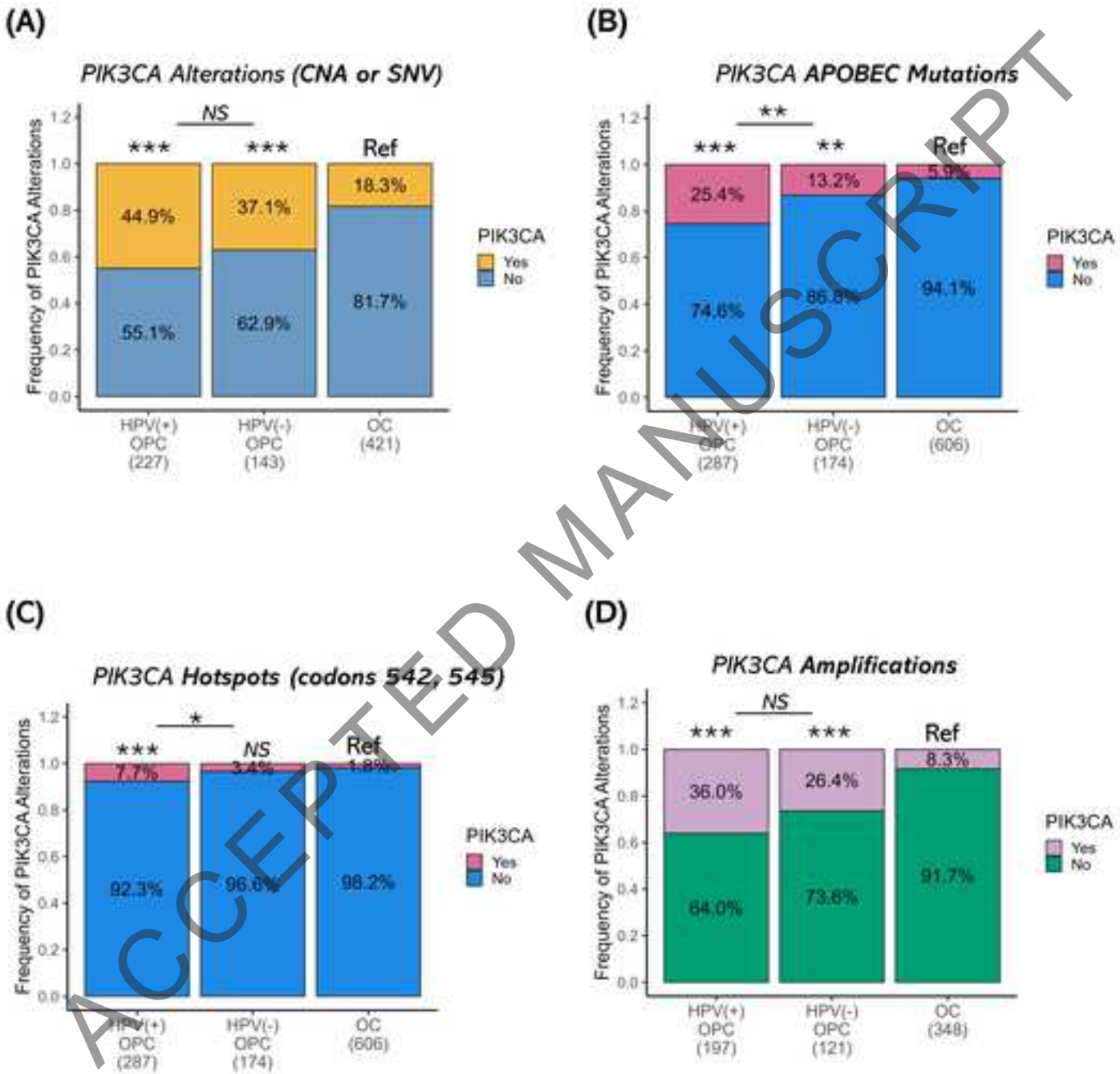


ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 2. Distribution of PIK3CA alterations by cancer subsite. Click here to access/download;Figure - Color;Figure2.tif

Figure 2



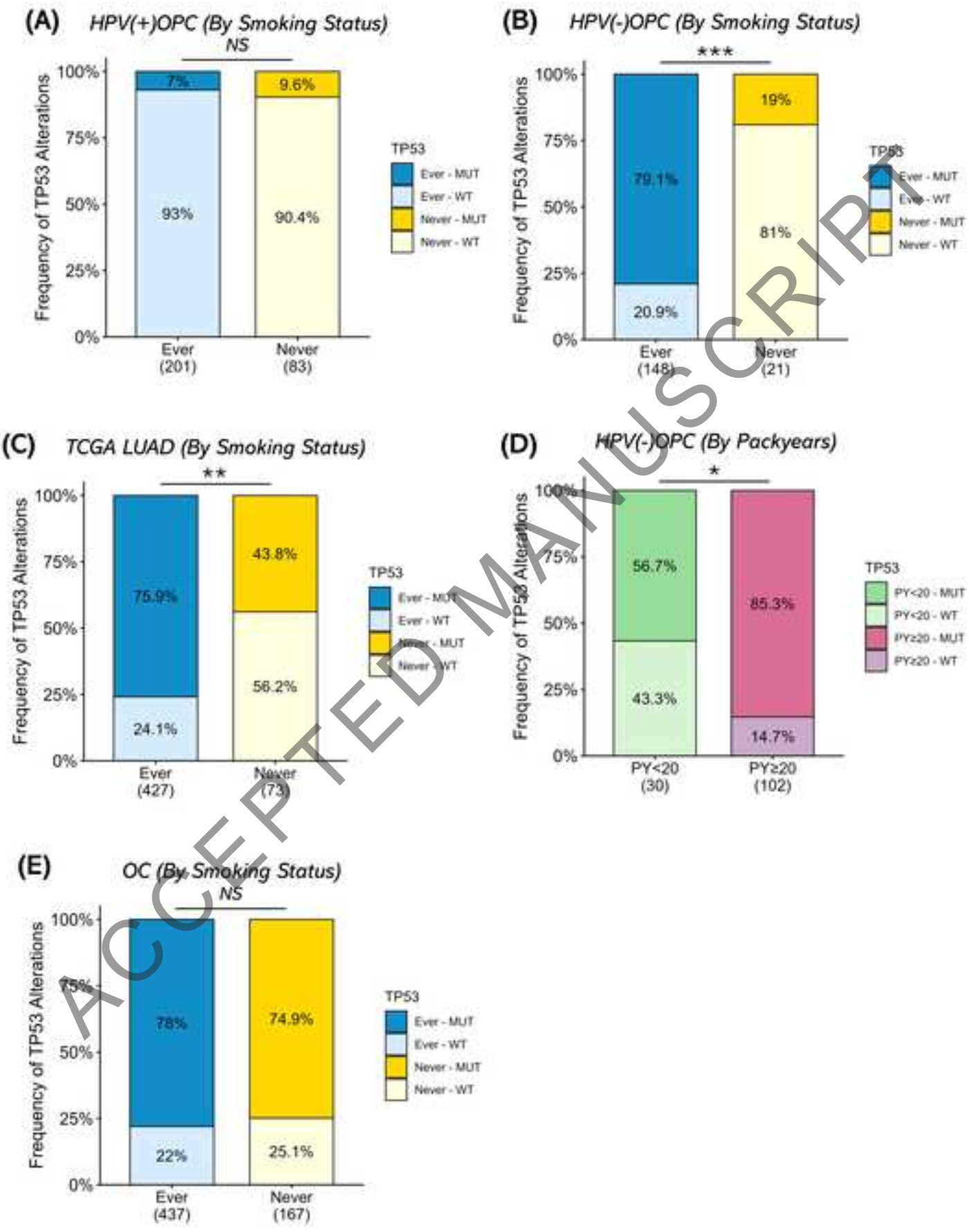
ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 3. Distribution of TP53 alterations by smoking status and intensity across tumor types.

Click here to access/download;Figure - Color;Figure3.tif

Figure 3



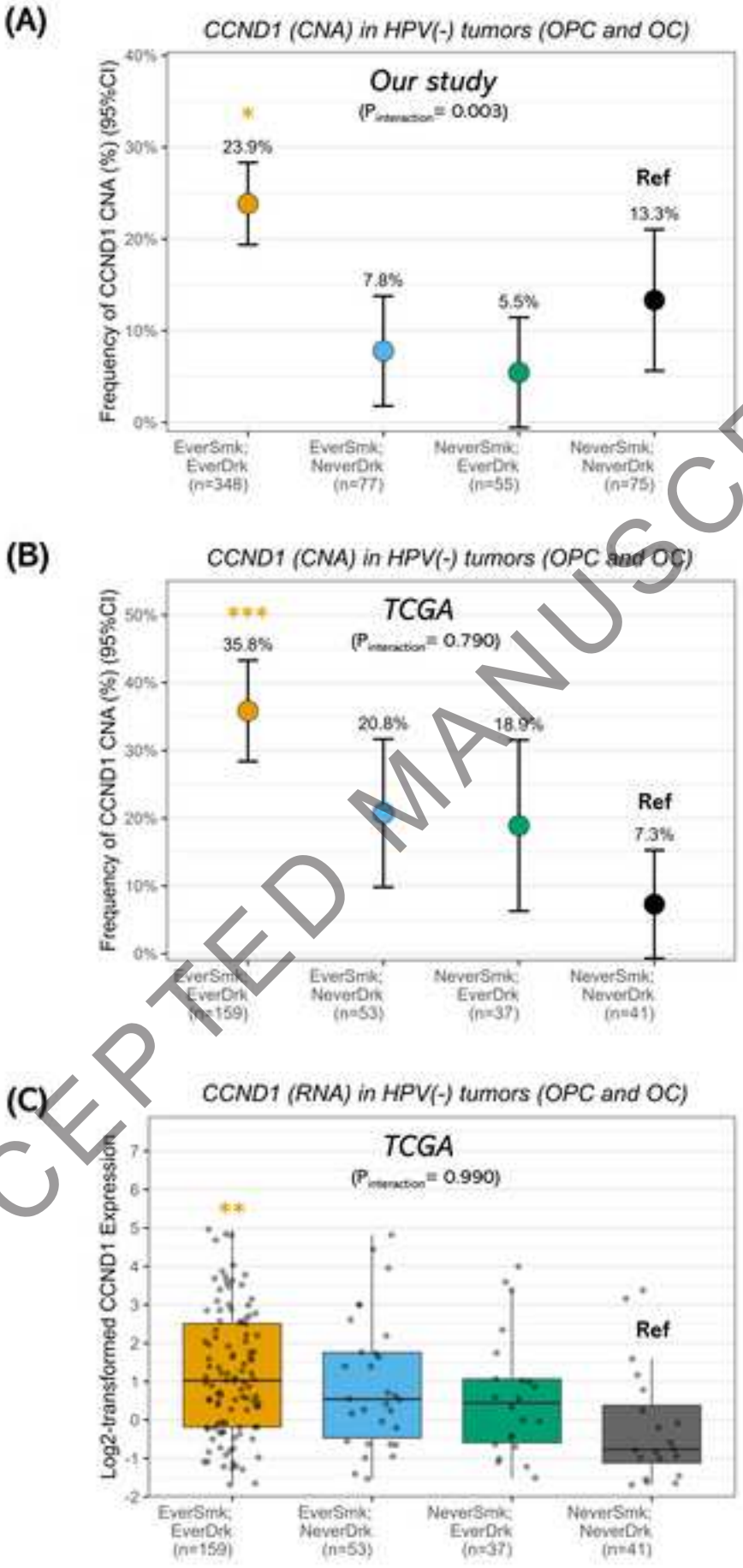
ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 4. CCND1 amplifications and expression by tobacco and alcohol exposure in HPV-negative tumors.

Click here to access/download;Figure - Color;Figure4.tif

Figure 4



ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 5. Distribution of the driver genes between oral cavity (OC) and oropharyngeal (OPC) anatomical sites. Heatmap representing

Click here to access/download;Figure - Color;Figure5.tif

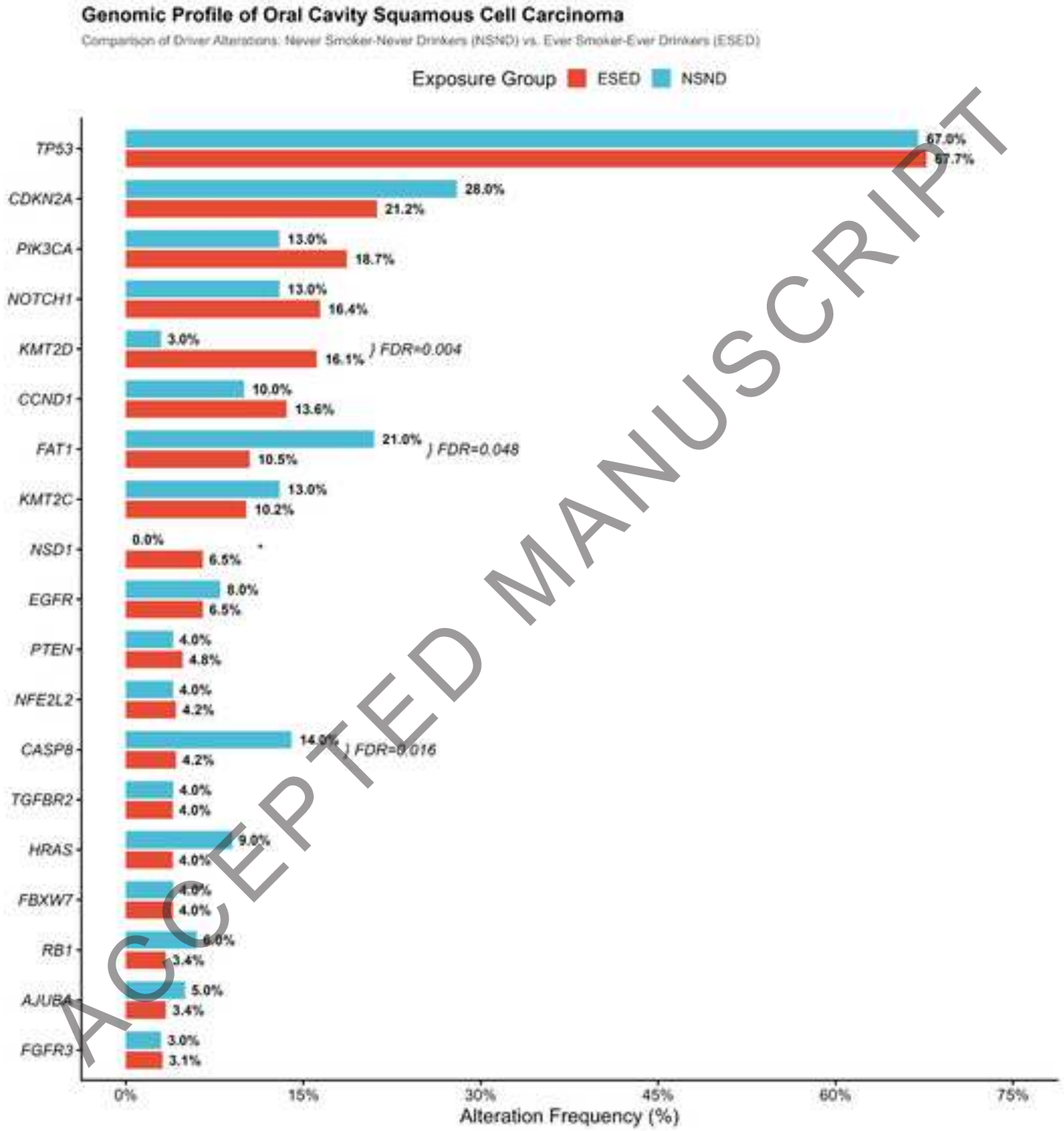
Figure 5

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026

Figure 6. Genomic profile of NSND OC tumors. Click here to access/download;Figure - Color;Figure6.tif

Figure 6



ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djag267/8760948 by INACTIVE user on 17 August 2026