

1 **The Complexity of Tobacco Smoke-Induced Mutagenesis in Head and Neck Cancer**

2

3 Laura Torrens¹, Sarah Moody², Ana Carolina de Carvalho¹, Mariya Kazachkova^{3,4,5}, Behnoush Abedi-
4 Ardekani¹, Saamin Cheema², Sergey Senkin¹, Thomas Cattiaux¹, Ricardo Cortez Cardoso Penha¹,
5 Joshua R. Atkins⁶, Valérie Gaborieau¹, Priscilia Chopard¹, Christine Carreira⁷, Ammal Abbasi^{3,8,5}, Erik
6 N. Bergstrom^{3,5}, Raviteja Vangara^{3,5}, Jingwei Wang², Stephen Fitzgerald², Calli Latimer², Marcos Diaz-
7 Gay^{3,5}, David Jones², Jon Teague², Felipe Ribeiro Pinto⁹, Luiz Paulo Kowalski¹⁰, Jerry Polesel¹¹, Fabiola
8 Giudici¹¹, José Carlos de Oliveira¹², Pagona Lagiou¹³, Areti Lagiou¹³, Marta Vilensky¹⁴, Dana Mates¹⁵,
9 Ioan N. Mates^{16,17}, Lidia M. Arantes¹⁸, Rui Reis¹⁸, Jose Roberto V. Podesta¹⁹, Sandra V. von Zeidler²⁰,
10 Ivana Holcatova²¹, Maria Paula Curado²², Cristina Canova²³, Elenora Fabianova²⁴, Paula A. Rodríguez-
11 Urrego²⁵, Laura Humphreys², Ludmil B. Alexandrov^{3,26,5,27}, Paul Brennan¹, Michael R. Stratton²,
12 Sandra Perdomo^{1*}

13

14 ¹Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon,
15 France, ²Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Cambridge, UK,
16 ³Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, USA,
17 ⁴Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, USA, ⁵Moores
18 Cancer Center, University of California San Diego, La Jolla, USA, ⁶Cancer Epidemiology Unit, The
19 Nuffield Department of Population Health, University of Oxford, Oxford, UK, ⁷Evidence Synthesis and
20 Classification Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France,
21 ⁸Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla,
22 USA, ⁹Brazilian National Cancer Institute, Rio de Janeiro, Brazil, ¹⁰University of São Paulo Medical
23 School, São Paulo, Brazil, ¹¹Unit of Cancer Epidemiology, Centro di Riferimento Oncologico di Aviano
24 (CRO) IRCCS, Aviano, Italy, ¹²Associação de Combate ao Câncer em Goiás Hospital Araújo Jorge,
25 Goiânia, Brazil, ¹³School of Medicine, National and Kapodistrian University of Athens, Athens,
26 Greece, ¹⁴Instituto de Oncología “Angel Roffo” Universidad de Buenos Aires, Buenos Aires,

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

27 Argentina, ¹⁵National Institute of Public Health, Bucharest, Romania, ¹⁶Carol Davila University of
28 Medicine and Pharmacy, Bucharest, Romania, ¹⁷Saint Mary Clinic of General and Esophageal Surgery,
29 Bucharest, Romania, ¹⁸Barretos Cancer Hospital, Barretos, Brazil, ¹⁹Hospital Santa Rita de Cássia -
30 Associação Feminina de Educação e Combate ao Câncer (AFECC), Vitória, Brazil, ²⁰Pathology
31 Department, Federal University of Espírito Santo, Vitória, Brazil, ²¹Charles University in Prague, 2nd
32 Faculty of Medicine, IHPM, Prague, Czech Republic, ²²A.C.Camargo Cancer Center, São Paulo, Brazil,
33 ²³Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thoraco-Vascular
34 Sciences and Public Health, University of Padua, Padova, Italy, ²⁴Regional Authority of Public Health,
35 Slovak Republic, ²⁵Hospital Universitario Fundación Santa Fe de Bogotá, Bogotá, Colombia,
36 ²⁶Department of Bioengineering, University of California San Diego, La Jolla, USA, ²⁷Sanford Stem Cell
37 Institute, University of California San Diego, La Jolla, USA
38 * Corresponding author: Sandra Perdomo

39 **ABSTRACT**

40 Tobacco smoke, alone or combined with alcohol, is the predominant cause of head and neck cancer
41 (HNC). Here, we further explore how tobacco exposure contributes to cancer development by
42 mutational signature analysis of 265 whole-genome sequenced HNC from eight countries. Six
43 tobacco-associated mutational signatures were detected, including some not previously reported.
44 Differences in HNC incidence between countries corresponded with differences in mutation burdens
45 of tobacco-associated signatures, consistent with the dominant role of tobacco in HNC causation.
46 Differences were found in the burden of tobacco-associated signatures between anatomical
47 subsites, suggesting that tissue-specific factors modulate mutagenesis. We identified an association
48 between tobacco smoking and three additional alcohol-related signatures indicating synergism
49 between the two exposures. Tobacco smoking was associated with differences in the mutational
50 spectra and repertoire of driver mutations in cancer genes, and in patterns of copy number change.
51 Together, the results demonstrate the multiple pathways by which tobacco smoke can influence the
52 evolution of cancer cell clones.

53 INTRODUCTION

54 Head and neck cancer (HNC), including malignancies affecting the mouth, pharynx, and larynx,
55 represents ~4% of the global cancer burden, with an annual incidence of about 750,000 new cases¹.

56 The incidence rate of HNC varies between different countries, largely reflecting the distribution of its
57 main risk factors including tobacco smoking, alcohol consumption^{1,2}, and infection with high-risk
58 strains of human papillomavirus (HPV) for oropharynx cancer³⁻⁵. Other proposed risk factors include
59 consumption of hot beverages, obesity, and poor oral health, although evidence for their role in HNC
60 is limited⁶⁻⁸. In addition, a substantial proportion of head and neck cancers (about 42% for women
61 and 26% for men) cannot be attributed to known lifestyle habits or exposures⁹.

62

63 Epidemiological studies in Europe and America suggest that seven out of 10 HNC cancers are caused
64 by preventable behavioral risk factors, with tobacco use, either alone or in combination with alcohol,
65 accounting for most cases⁹. Conversely, alcohol use on its own is responsible for only ~4% of the
66 disease burden, suggesting a limited impact on HNC burden. This raises the question of whether
67 alcohol acts as an independent carcinogen or simply enhances the known carcinogenic effect of
68 tobacco. Furthermore, the susceptibility to these exposures varies depending on the anatomical
69 region, with smoking posing a higher risk for developing larynx cancer and the risk associated with
70 alcohol being greater for other subsites¹⁰.

71

72 Considering the dominant role of tobacco in HNC development, risk differences across subsites, and
73 potential interactions with other risk factors, HNC offers a particularly interesting opportunity to
74 investigate the effects of tobacco exposure. In this context, the analysis of mutational signatures is
75 an effective tool to track the complex mutagenic patterns linked to this and other exposures over a
76 patient's lifetime¹¹⁻¹³. Certain mutational signatures have been related to well-established biological
77 mechanisms and exposures. Signatures SBS4, found predominantly in lung cancer, and SBS92, in

78 bladder cancer, capture two distinct mutagenic processes linked to tobacco use^{12,14,15}. Conversely,
79 Signature SBS16 has been attributed to alcohol consumption in esophageal and liver cancer^{13,16}.

80

81 Previous studies exploring the genomic landscape of HNC have relied predominantly on exome
82 sequencing data, which has limited power to detect mutational signatures, lacked a diverse
83 geographical and ethnic representation of cases, and/or were limited to specific anatomical
84 subsites¹⁷⁻²⁰. Therefore, the carcinogenic mechanisms underpinning this cancer type in different
85 geographical regions and anatomical subsites remain unclear. To bridge this gap, we performed
86 whole-genome sequencing of 265 HNC cases from individuals exposed to known and suspected risk
87 factors across eight countries with varying incidence rates. By leveraging mutational signature
88 analysis combined with extensive epidemiological data, we shed light on the complexity of tobacco-
89 induced mutagenesis and its interplay with alcohol consumption and other HNC risk factors.

90 RESULTS

91 Case-series overview and multi-country study design

92 A total of 265 HNC cases were included in the study, comprising retrospective collections from eight
93 countries in Europe and South America^{6,21} (**Figure 1; Supplementary Table 1**). These encompass a
94 broad geographic representation of HNC, including cases from high-incidence regions, with sex-
95 combined age-standardized rates (ASR) ranging from 9.4 per 100,000 to 18.2 per 100,000 in
96 Romania, Slovakia, Czech Republic, and Brazil, as well as moderate-incidence regions, with ASR from
97 3.8 to 7.8 per 100,000, in Colombia, Argentina, Greece, and Italy¹. The dataset contains cases from
98 all HNC anatomical subsites, with 127 oral cavity, 46 oropharynx, 17 hypopharynx, and 75 larynx
99 cancers. Epidemiological questionnaire data were available on exposure to known and suspected
100 HNC risk factors, including cases from drinkers, smokers, with both exposures, and non-exposed.
101 DNAs from paired tumor and blood samples were extracted and whole-genome sequenced to
102 average coverage of 55-fold and 27-fold, respectively.

103

104 Mutation burden

105 Among the 265 HNC cases, we observed a median of 12,887 single-base substitutions (SBS, range:
106 720 to 244,026), 63 doublet-base substitutions (DBS, range: two to 7,113), and 757 small insertions
107 and deletions (ID, range: 124 to 9,898; **Supplementary Table 2**). Tumor samples from tobacco users
108 exhibited higher SBS, DBS, and ID burdens compared to non-smokers (**Extended Data Figure 1b;**
109 **Supplementary Table 3**), as previously reported for larynx cancer¹⁴. Differences were also found
110 between anatomical subsites, with larynx samples presenting higher mutation burdens, even after
111 correcting for tobacco status (**Extended Data Figure 1a; Supplementary Table 3**). No significant
112 differences were found between geographical regions (**Extended Data Figure 1c**).

113

114 **Mutational signatures of exogenous and endogenous exposures**

115 To investigate the mutational processes and carcinogenic exposures that have been operative in
116 HNC development, we extracted SBS, DBS, and ID signatures and estimated the contribution of each
117 signature to every sample. We obtained 15 *de novo* SBS signatures, which were decomposed into 18
118 reference signatures from the Catalogue of Somatic Mutations in Cancer (COSMICv3.2) database,
119 and two signatures that could not be decomposed into any combination of existing signatures, SBS_I
120 and SBS_L (**Figure 2a-b; Extended Data Figure 2; Supplementary Tables 4,7-9, Supplementary**
121 **Note**).

122

123 Among the identified signatures, several have been previously associated with exogenous
124 mutational processes¹². The tobacco-related signatures SBS4 and SBS92 were found in 34% and 7.6%
125 of HNC samples and respectively accounted for 6.3% and 3.5% of the mutational burden on average.
126 SBS16, attributed to alcohol consumption^{13,16}, was present in 19% of the samples with a modest
127 impact on the HNC mutation burden of 1.4% on average. Signatures of ultraviolet (UV) light
128 exposure SBS7a and SBS7b co-occurred in 4.2% of cases.

129

130 We also identified signatures associated with endogenous exposures and aberrant cellular
131 processes. Notably, SBS2 and SBS13, which result from cytosine deamination by Apolipoprotein B
132 mRNA-editing enzyme catalytic polypeptide-like (APOBEC)¹⁴, were present in the majority of HNC
133 cases (93% and 92%, respectively; **Figure 2a**) and were highly correlated (**Supplementary Figure 1**).
134 Combined, these signatures accounted for an average of 20.4% of the total SBS mutation burden.
135 Other prevalent signatures included SBS18, which is caused by reactive oxygen species (77% of
136 samples), and clock-like signatures SBS1 (78%) and SBS5 (55%) (**Figure 2a**).

137

138 Extraction of DBS signatures identified four *de novo* signatures, which decomposed into four COSMIC
139 reference signatures (DBS1, DBS2, DBS4, and DBS6) and one non-decomposed signature (DBS_D;

140 **Figure 2a-b; Extended Data Figure 3a; Supplementary Tables 5,7-9).** We also extracted seven *de*
141 *novo* ID signatures, all of which were decomposed into 12 COSMIC signatures (**Figure 2a-b; Extended**
142 **Data Figure 3b; Supplementary Tables 6-9).** DBS and ID signatures of exogenous exposures were
143 positively correlated with their SBS counterparts (**Supplementary Figure 1**). For instance, the known
144 tobacco-related signatures DBS2 (59% of samples) and ID3 (41%), along with DBS6 which has been
145 previously registered as of unknown etiology, correlated with both SBS4 and SBS92. These
146 associations are consistent with these SBS, DBS, and ID signatures being generated by the same
147 underlying mutational process. Similarly, ID11 (38%), which was associated with alcohol
148 consumption in esophageal cancer¹³, exhibited a positive correlation with the alcohol signature
149 SBS16, while UV-related DBS1 (16.6%) and ID13 (1.5%) signatures showed the same link with SBS7a-
150 c.

151

152 To establish which mutagenic exposures were active earlier or later during the development of HNC,
153 we estimated the molecular timing of each SBS signature (**Methods**). Signatures of tobacco and
154 alcohol consumption, as well as the SBS_L signature, were enriched in early clonal mutations
155 (**Extended data Figure 4**), consistent with carcinogenic exposures occurring in normal cells²².
156 Similarly, SBS_I was significantly enriched in early clonal mutations in oral cavity cases, while no
157 significant differences were seen in other subsites. Signatures of APOBEC signaling and SBS39 were
158 enriched in late clonal mutations, suggesting that the corresponding mutational processes increased
159 in activity during the evolution of cancer clones²².

160

161 **HNC tumors present complex tobacco-related mutation patterns**

162 Several signatures were independently associated with tobacco consumption, including the
163 previously-recognized tobacco-related signatures SBS4, SBS92, DBS2, and ID3, as well as signature
164 DBS6, reported as of unknown etiology, and the newly-discovered SBS_I (**Figure 3a-b;**
165 **Supplementary Table 10; Supplementary Note**). The tobacco-associated SBS signatures were

166 composed of three different substitution patterns (predominantly C>A for SBS4, T>C for SBS92, and
167 T>A for SBS_I; **Figure 2c**) and exhibited transcriptional strand bias (**Supplementary Figure 4**)^{15,23}. This
168 strand bias towards the transcribed strand often occurs as a result of transcription-coupled DNA
169 repair and is found in mutations due to bulky adducts, caused by exogenous exposures such as
170 tobacco smoke carcinogens²³. Assuming this mechanism is responsible for the strand bias in SBS_I,
171 this is indicative of adduct formation on adenine bases.

172
173 The distribution of tobacco-associated signatures varied across different anatomical subsites (**Figure**
174 **3a, Extended Data Figure 5a; Supplementary Table 10**). Previously established tobacco signatures
175 exhibited higher signature burdens and frequencies in larynx cases compared to other subsites. For
176 instance, SBS4 was present in 17% of OC, 17% of oropharynx, 53% of hypopharynx, and 67% of
177 larynx cases. Similar distributions were observed for SBS92, DBS2, and ID3. Conversely, the
178 previously unknown SBS_I signature was present in smokers across all subsites, with particular
179 enrichment in the oral cavity. The associations between signatures and subsites remained significant
180 after correction for tobacco consumption and other confounding variables (**Supplementary Table**
181 **10**).

182

183 **Effects of tobacco exposure on the driver mutation spectra**

184 We explored the driver mutation profile in tobacco-related HNC. This revealed 96 cancer genes with
185 driver mutations in our dataset, including *TP53*, *NOTCH1*, *CDKN2A*, *KMT2D*, and *CASP8*, which are
186 commonly implicated in HNC²⁴ (**Extended Data Figure 6a-b; Supplementary Tables 11-12**). *TP53*
187 mutations were significantly enriched among smokers compared to non-smokers (83% [164/197] vs
188 61% [42/68], $p=0.001$), while *CASP8* mutations were more frequent among non-smokers (6.09%
189 [12/197] vs 20.6% [14/68], $p=0.003$). A total of 642 driver mutations were identified (**Methods**), and
190 these showed an enrichment of C>A substitutions in smokers compared to non-smokers (24.9%
191 [114/457] vs 17.3% [32/185], Fisher's exact test $p=0.0379$; **Extended Data Figure 6c**), consistent with

192 the SBS4 mutation profile¹². The frequency of C>A driver mutations in tobacco-exposed cases was
193 higher in the larynx subsite compared to oral cavity (31.5% [53/168] vs 19.9% [38/191], Fisher's
194 exact test p=0.0148; **Figure 3c-d**). This reflects the lower contribution of SBS4 to mutations in
195 tobacco-exposed oral cavity HNC compared to larynx cases, which has been carried through into the
196 generation of driver mutations. T>A driver mutations were also observed among smokers, albeit in
197 low frequencies (6.6% [11/168] in larynx and 8.4% [16/191] in oral cavity, hinting at a lower
198 presence of SBS_I in driver mutations.

199

200 **Tobacco-related mutational signatures correlate with demographic HNC incidence**

201 We analyzed the link between tobacco mutagenesis and variations in HNC incidence across different
202 countries, sexes, and anatomical subtypes. Our findings support previous epidemiological evidence,
203 which has shown a connection between HNC incidence and smoking habits² (**Figure 4a-b**).
204 Moreover, HNC incidence correlated with tobacco-related signatures (**Figure 4c; Supplementary**
205 **Figure 2**), showing a higher ASR of HNC incidence in demographic groups presenting higher signature
206 burdens. This further confirms that the geographical and demographic differences in tobacco
207 exposure play a dominant role in driving HNC incidence.

208

209 **Alcohol-related mutational signatures in drinkers and smokers**

210 Next, we assessed the signature profile in HNC cases with a history of alcohol intake. Our analysis
211 revealed significant associations between alcohol consumption and three specific signatures: SBS16,
212 ID11, and DBS4 (**Figure 5; Supplementary Table 10; Supplementary Note**). SBS16 was present
213 exclusively in HNC cases from drinkers and showed enrichment in samples exposed to both tobacco
214 and alcohol compared to alcohol alone (29.0% [47/162] and 12.5% [4/32], respectively). Similarly,
215 DBS4 and ID11 also presented higher burdens and signature frequencies in cases exposed to both
216 risk factors (**Figure 5; Supplementary Table 10**). Although the etiology of DBS4 is unclear, it has been
217 found prevalent in esophageal cancer cases from countries with high alcohol intake rates¹³. The

218 results are, therefore, consistent with SBS16, DBS4 and ID16 all being generated by the same
219 underlying alcohol-related mutational process and that the mutagenicity of this process is increased
220 with co-exposure to tobacco smoke.

221

222 For driver mutations, samples from individuals exposed to both tobacco and alcohol were
223 characterized by a particularly high *TP53* frequency of mutations (87.0% [141/162], 71.4% [25/35],
224 68.8% [22/32], and 55.6% [20/36] in the tobacco plus alcohol, alcohol alone, tobacco alone, and
225 unexposed groups, respectively, Fisher's exact test $p=0.0001$; **Extended Data Figure 6**;
226 **Supplementary Table 11**). The driver mutation burden in the SBS16 context was too low to assess
227 differences in the driver spectra between groups. However, *TP53* mutations in the SBS16 contexts
228 were exclusively found in samples from individuals exposed to both tobacco and alcohol ($n=5$ *TP53*
229 variants).

230

231 **HPV-positive HNC is characterized by APOBEC signatures**

232 HPV infection in oropharynx cases did not elicit a specific mutational signature profile
233 (**Supplementary Table 10**). However, the substitution profile in HPV-positive oropharyngeal HNC
234 was characterized by a higher relative proportion of APOBEC signatures, with 57.6% of the signature
235 burden being attributed to SBS2 and SBS13 on average, compared to 30.0% in HPV-negative
236 oropharynx (**Extended Data Figure 7a**) consistent with previous reports¹⁸. Notably, the presence of
237 APOBEC signatures was nearly ubiquitous across HNC cases (**Figure 2a**), suggesting a broader role for
238 APOBEC activation beyond its anti-viral function¹¹.

239

240 We also observed differences between HPV-positive and HPV-negative oropharynx cases exposed to
241 tobacco. Among smokers, only 1/6 (17%) HPV-positive oropharynx cases presented tobacco-related
242 SBS signatures, compared to 7/26 (27%) in HPV-negative cases (Fisher's exact test $p=0.0214$).
243 Despite the well-known influence of tobacco smoking on the driver profiles of HNC^{20,24}, the driver

244 alterations in HPV-positive smokers differed from that of HPV-negative smokers, and instead
245 resembled the profile in HPV-positive cases from non-smokers. This included *PIK3CA* mutations,
246 *PTEN* mutations and deletions, as well as absence of *TP53* mutations and of *FADD* gains (**Extended**
247 **Data Figure 7b-c**). This, together with the reduced presence of tobacco-related signatures, suggests
248 that oncogenesis in HPV-positive smokers may primarily be driven by viral infection rather than
249 tobacco exposure.

250

251 **Mutational signature profile in samples exposed to putative HNC risk factors**

252 We next investigated the presence of additional environmental exposures beyond the most widely-
253 known HNC risk factors. Notably, UV-related signatures SBS7a-c, DBS1, and ID13 were detected
254 predominantly in oral cavity cases (**Figure 6a; Supplementary Table 10; Supplementary Note**). SBS7
255 signatures have been previously described in HNC, but the anatomical and epidemiological features
256 of positive cases have not been previously investigated¹². Samples with a relative SBS7a-c burden of
257 >10% were categorized as positive for UV exposure, a criterion met by 13 oral cavity cases from the
258 lip, tongue, and floor of the mouth (**Figure 6b**). All positive cases were either tobacco or alcohol
259 users, with 11/13 presenting both risk factors (**Figure 6b**). Thus, our data suggests a potential role of
260 UV light exposure in HNC carcinogenesis²³, which could be enhanced by tobacco and/or alcohol.

261

262 Our analysis did not show any specific mutational patterns associated with other putative HNC risk
263 factors, including hot drink consumption, poor oral health score, and high body mass index^{6,7}
264 (**Supplementary Table 10**). This suggests that these agents are likely not causing direct mutagenesis.
265 Finally, the previously unknown DBS_D signature and ID4, with unknown etiology, were enriched
266 among non-smokers (**Extended Data Figure 5b**), suggesting a potential link to unidentified
267 mutational processes in this population.

268

269 **HNC risk factors elicit distinct copy number profiles**

270 HNC is characterized by complex patterns of copy number (CN) aberrations throughout the
271 genome^{19,20}. Unsupervised hierarchical clustering analysis on the CN counts in HNC samples ($n=242$)
272 revealed two main clusters, one displaying diploid genomes (cluster D), and another presenting
273 polyploidy and high burden of CN gains and losses (cluster P; **Extended Data Figure 8**;
274 **Supplementary Figure 3**). These clusters further subdivided into four groups (D1, D2, P1, and P2).
275 Notably, subgroup D2 was characterized by a CN-neutral profile, exhibiting significantly lower
276 burdens of CN events compared to the other groups.

277
278 The CN clusters were associated with distinct epidemiological profiles (**Figure 7c-d**; **Supplementary**
279 **Table 13**). Specifically, tobacco-related HNC were enriched within both the diploid and polyploid CN-
280 high clusters (i.e., D1, P1, and P2), while the CN-silent cluster D2 was mostly constituted by samples
281 from non-smokers, including cases with unknown risk factors and alcohol drinkers in the absence of
282 tobacco. Consistent with this pattern, the D2 cluster was enriched in samples from female patients,
283 oral cavity cases, and older age, aligning with the characteristic features of HNC with undefined risk
284 factor²⁴. Finally, HPV-positive oropharynx cases were enriched in the diploid clusters, predominantly
285 in cluster D1.

286
287 To unveil distinct CN particularities within each CN cluster and etiology, we conducted CN signature
288 analysis²⁵ (**Figure 7a-b**; **Extended Data Figure 9**; **Supplementary Note**). Cluster D1 exhibited
289 enrichment in signatures of chromosomal instability within a diploid genome background (signatures
290 CN1, CN9 and CN13; **Extended Data Figure 10a**). In contrast, cluster D2 presented a signature profile
291 related to a diploid copy-neutral background (CN1). Clusters P1 and P2 displayed associations with
292 signatures of whole-genome duplication (CN2, CN20) along with genomic aberrations (CN5, CN_G;
293 **Extended Data Figure 10b-c**). Cluster P1 was consistent with double whole-genome duplications
294 (CN18), while P2 showed signatures of chromosomal instability in conjunction with genome doubling

295 (CN12). Collectively, our analysis suggests that HNC risk factors align with different CN profiles and
296 provides an enhanced characterization of the CN aberrations in each HNC etiology (**Figure 7d**).
297 Specifically, tobacco use, alone or with alcohol, may trigger chromosomal instability and aneuploidy,
298 while HPV infection may confer a CN unstable diploid profile. Lastly, samples with unknown risk
299 factors exhibit a CN-neutral profile.

300

301 We explored whether this difference in the CN profile could be due to the driver profile that is
302 associated with each risk factor (**Figure 7d; Extended Data Figure 10d**). *TP53* mutations and *MYC*
303 gains, two known promoters of genomic instability²⁵⁻²⁷, as well as gains in the anti-apoptotic *FADD*
304 gene, were enriched in cluster P. *CASP8* and *HRAS* mutations were enriched in the D2 CN-neutral
305 cluster, in agreement with previous studies in HNC^{20,24,25}. Finally, *PTEN* and *RB1* mutations were
306 enriched in the D1 cluster. Overall, these results show that tobacco use in HNC is associated with a
307 distinct CN-rich profile and driver alterations related to genome instability.

308 DISCUSSION

309 The role of tobacco as one of the most avoidable cancer risk factors has been known for over 50
310 years. Yet, the detailed mechanisms by which tobacco smoke leads to DNA damage and
311 carcinogenesis in different tissues are still not fully understood^{14,28,29}. In this study encompassing
312 HNC cases from eight countries in Europe and South America, we shed light on the effects of tobacco
313 as the main mutagenic exposure in HNC and explored the complex mutational patterns and genomic
314 alterations linked to tobacco exposure in different HNC subsites, as well as its interplay with alcohol
315 consumption and other risk factors.

316

317 Tobacco smoke contains a mixture of thousands of chemicals, including over 60 carcinogens, among
318 which benzo[*a*]pyrenes (BaP) and nitrosamines are the most widely studied. These carcinogens
319 undergo metabolic activation, generating reactive intermediates that interact with DNA in exposed
320 tissues, resulting in complex mutagenic processes that can lead to cancer development²⁹. In HNC,
321 tobacco exposure resulted in six different signatures, unveiling at least three mutational processes
322 due to tobacco in HNC. Signature SBS4, characterized by C>A transversions, has been largely
323 attributed to BaP adducts^{14,30,31}. Exposure to this compound is also consistent with the CC>AA
324 substitutions and C deletions present in DBS2 and ID3 tobacco signatures, respectively³¹. Conversely,
325 signature SBS92, composed predominantly of T>C transitions, has not been related to specific
326 carcinogens in tobacco smoke¹⁵. Finally, the T>A-rich substitution profile captured by the previously-
327 unidentified signature SBS_I is compatible with adduct formation on adenines, which have been
328 observed in response to multiple tobacco compounds³¹⁻³³. Among those, exposure to nicotine-
329 derived nitrosamine ketone (NNK), one of the main tobacco carcinogens in oral tissues³⁴, also
330 yielded a T>A-rich signature *in vitro*^{35,36}. Notably, a signature exhibiting high T>A frequencies and
331 transcriptional strand bias has been described in normal lung epithelia from patients with a history
332 of smoking³⁷.

333

334 Our epidemiological analysis revealed that the mutational effects of tobacco vary among anatomical
335 subsites. The canonical tobacco signatures SBS4 and SBS92 were found predominantly in larynx
336 cases, along with the tobacco-related DBS and ID signatures. Conversely, SBS_I was extracted in HNC
337 cases from all subsites, with a notable enrichment in oral cavity cases. Altogether, our observations
338 hint at varying susceptibility, exposure level, or clearance of tobacco carcinogens across tissues,
339 leading to different genotoxic effects. A possible explanation for these differences is the tissue-
340 specific pattern of cytochrome P450 function. CYP1A1, the main BaP metabolizer, is primarily
341 expressed in lung and larynx, whereas enzymes responsible for nitrosamine metabolism, such as
342 CYP2E1, are predominant in the upper aerodigestive tract including the oral cavity^{34,38-40}. These
343 differences in the response to tobacco across tissues may partially explain the greater susceptibility
344 to smoking found for larynx cancers compared to other anatomical subsites¹⁰. While tobacco use
345 was associated with elevated mutation burdens and BaP-related driver mutations in larynx cancers,
346 this was not observed in oral cavity cases, aligning with a reduced carcinogenic effect. Thus,
347 additional carcinogenic processes may be necessary to aid in the development of oral cavity and
348 oropharynx cancers, including alcohol and HPV infection¹⁰.

349
350 In this regard, we also identified mutagenic processes linked to alcohol exposure^{13,16}, including
351 signatures SBS16, ID11, and an unreported association with signature DBS4. In HNC, alcohol-related
352 signatures were predominantly observed in patients reporting both alcohol and tobacco
353 consumption, reflecting the synergistic effect between these two factors on disease risk^{9,41}.
354 Furthermore, a previous study suggested an enrichment of SBS16 in oropharynx cases from tobacco
355 users¹⁸. Altogether, our findings indicate that tobacco could enhance the carcinogenic effects of
356 alcohol through shared mutagenic processes. Experimental evidence suggests that salivary
357 concentrations of acetaldehyde, the genotoxic byproduct of alcohol metabolism, are greatly
358 increased by tobacco smoking⁴², which could result in enhanced alcohol-related mutagenesis in
359 cases with combined exposure.

360

361 Our data show that tobacco use, alone or in conjunction with alcohol, is also associated with a
362 distinct CN-rich profile, characterized by CN signatures of chromosomal instability, and resembling a
363 previously described subset of CN-rich HNC¹⁹. These genomic profiles are likely due to driver
364 alterations leading to genome instability such as *TP53* mutations, which are prevalent among
365 smokers and drinkers²⁴. Although high CN burdens have been reported in lung adenocarcinoma
366 cases from smokers¹⁴, the link between this exposure and specific CN or driver profiles in HNC was
367 previously unclear⁴³. Cases with unknown etiology, on the other hand, exhibit few CN alterations,
368 prevalence of *CASP8* and *HRAS* mutations, and wild-type *TP53*. A similar CN-neutral group of
369 samples has been observed in HNC, with an unreported link with HNC etiology^{19,44}.

370

371 Regarding the mutagenic potential of other investigated risk factors, HPV infection did not elicit a
372 specific mutational signature profile, but it was associated with distinct driver mutations and a CN-
373 unstable diploid genome. Poor oral hygiene, high body mass index, and consumption of hot drinks
374 did not display a direct effect on the mutation profile of HNC cases and likely contribute to the
375 development of HNC of unknown etiology through mechanisms distinct from direct mutagenesis.
376 This pattern has been proposed for several carcinogens in prior studies^{13,45}. Nevertheless, there may
377 exist additional unidentified mutagens leading to HNC, as hinted by the presence of the previously
378 unidentified signature SBS_L as well as the enrichment of DBS_D and ID4 among non-smokers.

379

380 Furthermore, we provide evidence suggesting that sunlight exposure may contribute to HNC
381 development. Specifically, we identified signatures consistent with pyrimidine dimer formation
382 (SBS7a-c and DBS1) in oral HNC cases, indicative of DNA damage by UV light^{12,23}. UV light has only
383 been described as a risk factor for malignancies in the external lip⁴⁶, but experimental evidence
384 suggests that oral cavity epithelia are susceptible to this exposure, and its carcinogenic processes
385 could be enhanced by tobacco smoking⁴⁷⁻⁵¹. While we cannot exclude the possibility of other

386 mutational processes eliciting CC>TT substitutions, such as those driven by reactive oxygen species⁵²,
387 the presence of ID13 signatures, identified in melanoma¹², provides additional evidence supporting
388 the role of sunlight exposure in oral HNC.

389

390 In summary, through our comprehensive analysis of the mutational, genomic, and epidemiological
391 profile of HNC cases from diverse geographical regions, we have uncovered genomic mechanisms by
392 which tobacco smoke and other risk factors contribute to HNC development. These findings enhance
393 our understanding of the complexity and tissue-specificity of tobacco mutagenesis, offering
394 additional evidence that may inform prevention strategies aimed at reducing the risk of this disease.

395 **REFERENCES**

- 396 1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality
397 worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **71**, 209–249 (2021).
- 398 2. Simard, E. P., Torre, L. A. & Jemal, A. International trends in head and neck cancer incidence
399 rates: differences by country, sex and anatomic site. *Oral Oncol* **50**, 387–403 (2014).
- 400 3. *Alcohol Consumption and Ethyl Carbamate. Alcohol Consumption and Ethyl Carbamate*
401 (2010).
- 402 4. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans., World Health
403 Organization. & International Agency for Research on Cancer. *Tobacco Smoke and Involuntary*
404 *Smoking*. (IARC Press, 2004).
- 405 5. International Agency for Research on Cancer, (IARC) & World Health Organization, (WHO).
406 *Human Papillomaviruses. Human papillomaviruses* vol. 90 (2007).
- 407 6. Hashim, D. *et al.* The role of oral hygiene in head and neck cancer: results from International
408 Head and Neck Cancer Epidemiology (INHANCE) consortium. *Annals of Oncology* **27**, 1619
409 (2016).
- 410 7. Gaudet, M. M. *et al.* Body mass index and risk of head and neck cancer in a pooled analysis of
411 case–control studies in the International Head and Neck Cancer Epidemiology (INHANCE)
412 Consortium. *Int J Epidemiol* **39**, 1091–1102 (2010).
- 413 8. Loomis, D. *et al.* Carcinogenicity of drinking coffee, mate, and very hot beverages. *Lancet*
414 *Oncol* **17**, 877–878 (2016).
- 415 9. Hashibe, M. *et al.* Interaction between tobacco and alcohol use and the risk of head and neck
416 cancer: Pooled analysis in the international head and neck cancer Epidemiology consortium.
417 *Cancer Epidemiology Biomarkers and Prevention* **18**, 541–550 (2009).
- 418 10. Lubin, J. H. *et al.* Total Exposure and Exposure Rate Effects for Alcohol and Smoking and Risk
419 of Head and Neck Cancer: A Pooled Analysis of Case-Control Studies. *Am J Epidemiol* **170**,
420 937–947 (2009).

- 421 11. Perdomo, S. *et al.* The Mutographs biorepository: A unique genomic resource to study cancer
422 around the world. *Cell Genomics* 100500 (2024) doi:10.1016/J.XGEN.2024.100500.
- 423 12. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**,
424 94–101 (2020).
- 425 13. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight
426 countries with varying incidence. *Nat Genet* **53**, 1553–1563 (2021).
- 427 14. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human
428 cancer. *Science* **354**, 618–622 (2016).
- 429 15. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with
430 SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
- 431 16. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and
432 cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315 (2017).
- 433 17. Plath, M. *et al.* Unraveling most abundant mutational signatures in head and neck cancer. *Int*
434 *J Cancer* **148**, 115–127 (2021).
- 435 18. Gillison, M. L. *et al.* Human papillomavirus and the landscape of secondary genetic alterations
436 in oral cancers. *Genome Res* **29**, 1–17 (2019).
- 437 19. Yang, J., Chen, Y., Luo, H. & Cai, H. The Landscape of Somatic Copy Number Alterations in
438 Head and Neck Squamous Cell Carcinoma. *Front Oncol* **10**, 479033 (2020).
- 439 20. Sayáns, M. P. *et al.* Comprehensive Genomic Review of TCGA Head and Neck Squamous Cell
440 Carcinomas (HNSCC). *Journal of Clinical Medicine* 2019, Vol. 8, Page 1896 **8**, 1896 (2019).
- 441 21. Slot, D. E., Van Der Weijden, F. & Ciancio, S. G. Oral health, dental care and mouthwash
442 associated with upper aerodigestive tract cancer risk in Europe: the ARCAGE study. *Oral*
443 *Oncol* **50**, e57 (2014).
- 444 22. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* 2020 578:7793 **578**,
445 122–128 (2020).

- 446 23. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–
447 421 (2013).
- 448 24. Leemans, C. R., Snijders, P. J. F. & Brakenhoff, R. H. The molecular landscape of head and neck
449 cancer. *Nat Rev Cancer* **18**, 269–282 (2018).
- 450 25. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **2022**
451 *606:7916* **606**, 984–991 (2022).
- 452 26. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nature Reviews Cancer* **2008 8:12** **8**,
453 976–990 (2008).
- 454 27. Steele, C. D., Pillay, N. & Alexandrov, L. B. An overview of mutational and copy number
455 signatures in human cancer. *J Pathol* **257**, 454 (2022).
- 456 28. Cogliano, V. J. *et al.* Preventable exposures associated with human cancers. *J Natl Cancer Inst*
457 **103**, 1827–1839 (2011).
- 458 29. Hecht, S. S. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature*
459 *Reviews Cancer* **2003 3:10** **3**, 733–744 (2003).
- 460 30. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis* **30**, 763
461 (2015).
- 462 31. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**,
463 821-836.e16 (2019).
- 464 32. Westcott, P. M. K. *et al.* The mutational landscapes of genetic and chemical models of Kras-
465 driven lung cancer. *Nature* **517**, 489 (2015).
- 466 33. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-
467 associated cancers. *Oncogene* **2002 21:48** **21**, 7435–7451 (2002).
- 468 34. Hecht, S. S. & Hatsukami, D. K. Smokeless Tobacco and Cigarette Smoking: Chemical
469 Mechanisms and Cancer Prevention. *Nat Rev Cancer* **22**, 143 (2022).
- 470 35. Peterson, L. A. Context matters: Contribution of specific DNA adducts to the genotoxic
471 properties of the tobacco-specific nitrosamine NNK. *Chem Res Toxicol* **30**, 420–433 (2017).

- 472 36. Mingard, C. *et al.* Dissection of Cancer Mutational Signatures with Individual Components of
473 Cigarette Smoking. *Chem Res Toxicol* **36**, 714–723 (2023).
- 474 37. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium.
475 *Nature* **578**, 266–272 (2020).
- 476 38. Degawa, M. *et al.* Metabolic activation and carcinogen-DNA adduct detection in human
477 larynx. *Cancer Res* **54**, 4915–9 (1994).
- 478 39. Jones, N. J., McGregor, A. D. & Waters, R. Detection of DNA adducts in human oral tissue:
479 correlation of adduct levels with tobacco smoking and differential enhancement of adducts
480 using the butanol extraction and nuclease P1 versions of 32P postlabeling. *Cancer Res* **53**,
481 1522–8 (1993).
- 482 40. Yamazaki, H., Inui, Y., Yun, C. H., Guengerich, F. P. & Shimada, T. Cytochrome P450 2E1 and
483 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and
484 tobacco-related nitrosamines in human liver microsomes. *Carcinogenesis* **13**, 1789–1794
485 (1992).
- 486 41. Hoes, L., Dok, R., Verstrepen, K. J. & Nuyts, S. Ethanol-Induced Cell Damage Can Result in the
487 Development of Oral Tumors. *Cancers 2021, Vol. 13, Page 3846* **13**, 3846 (2021).
- 488 42. Gapstur, S. M. *et al.* The IARC Perspective on Alcohol Reduction or Cessation and Cancer Risk.
489 (2023).
- 490 43. Pickering, C. R. *et al.* Squamous cell carcinoma of the oral tongue in young non-smokers is
491 genomically similar to tumors in older smokers. *Clin Cancer Res* **20**, 3842–3848 (2014).
- 492 44. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous
493 cell carcinomas. *Nature 2015 517:7536* **517**, 576–582 (2015).
- 494 45. Riva, L. *et al.* The mutational signature profile of known and suspected human carcinogens in
495 mice. *Nat Genet* **52**, 1189–1197 (2020).

- 496 46. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans., World Health
497 Organization. & International Agency for Research on Cancer. *Solar and Ultraviolet Radiation*.
498 (IARC, 1992).
- 499 47. Agrawal, A. *et al.* UV radiation increases carcinogenic risks for oral tissues compared to skin.
500 *Photochem Photobiol* **89**, 1193–1198 (2013).
- 501 48. von Koschembahr, A. *et al.* Solar simulated light exposure alters metabolism and
502 genotoxicity induced by benzo[a]pyrene in human skin. *Scientific Reports* **2018 8:1 8**, 1–12
503 (2018).
- 504 49. King, G. N. *et al.* Increased prevalence of dysplastic and malignant lip lesions in renal-
505 transplant recipients. *N Engl J Med* **332**, 1052–1057 (1995).
- 506 50. Sugano, N., Minegishi, T., Kawamoto, K. & Ito, K. Nicotine inhibits UV-induced activation of
507 the apoptotic pathway. *Toxicol Lett* **125**, 61–65 (2001).
- 508 51. Onoda, N. *et al.* Nicotine affects the signaling of the death pathway, reducing the response of
509 head and neck cancer cell lines to DNA damaging agents. *Head Neck* **23**, 860–870 (2001).
- 510 52. Reid, T. M. & Loeb, L. A. Tandem double CC-->TT mutations are produced by reactive oxygen
511 species. *Proc Natl Acad Sci U S A* **90**, 3904 (1993).
- 512
- 513

514 **FIGURE AND TABLE LEGENDS**

515 **Figure 1. Head and neck cancer incidence and epidemiological characteristics.** **a**, Incidence of head
516 and neck cancer (HNC), sex-combined, age-standardized rates (ASR) per 100,000, data from
517 GLOBOCAN 2022. Dots indicate countries included in this study and number of participating
518 patients. **b**, Anatomical subsites of HNC, with number of tumor samples indicated in brackets.
519 Created with biorender.com **c**, Known and suspected risk factors included in the study, based on
520 epidemiological questionnaire data and human papillomavirus (HPV) detection. Frequencies of risk
521 factors in the complete dataset (left) and by anatomical subsite (right) are indicated. OC, oral cavity;
522 OPC, oropharynx; HPX, hypopharynx; LYX, larynx.

523

524 **Figure 2. Mutational signature landscape of head and neck cancer.** **a**, Single base substitution (SBS),
525 doublet base substitution (DBS), and insertion deletion (ID) signatures extracted in 265 HNC tumors.
526 The size of each dot represents the proportion of samples presenting each mutational signature in
527 the whole HNC dataset and across anatomical subsites. The color represents the mean relative
528 attribution of each signature. Gray dots indicate signatures without significantly different relative
529 burdens by subsite. The top panel shows the mutations per megabase attributed to each signature
530 in samples with counts higher than zero. Significance was assessed using a two-sided Kruskal-Wallis
531 test and Bonferroni correction. **b**, Mutational spectrum of undecomposed signatures extracted from
532 HNC. **c**, Known signatures of tobacco exposure identified in the HNC dataset. HNC, head and neck
533 cancer; OC, oral cavity; OPC, oropharynx; ROS, reactive oxygen species; HR, homologous
534 recombination; DSB, double-strand break.

535

536 **Figure 3. Tobacco-related signatures.** **a**, Mutational burdens of tobacco-related signatures in HNC
537 cases sorted by subsite and tobacco status. Tumor mutational burdens (TMB) per sample is also
538 displayed. For clarity, y axis has been cut at 100,000 for TMB and at 15,000 for SBS_I. **b**, Mutational
539 burdens for SBS, DBS, and ID signatures showing significant positive associations with tobacco

540 consumption (n = 265 biologically independent samples). The Kruskal–Wallis test (two sided) was
541 used to test for global differences. Box-and-whisker plots are in the style of Tukey. The line within
542 the box is plotted at the median, while upper and lower ends indicate 25th and 75th percentiles.
543 Whiskers show 1.5 × interquartile range (IQR), and values outside it are shown as individual data
544 points. Y axes were cut at 1.25 × upper whisker for clarity. Bar plots indicate the frequencies of
545 dichotomized signatures. **c**, Percentage of driver mutations occurring in C>A contexts in larynx and
546 oral cavity HNC from smokers. **d**, SBS96-mutation spectrum of driver mutations in larynx and oral
547 cavity HNC from smokers, showing enrichment in the frequency of C>A driver mutations in larynx
548 cases. HPX, hypopharynx.

549
550 **Figure 4. Association of tobacco use with incidence of head and neck cancer.** **a**, Correlation
551 between age-standardized rate (ASR) of HNC incidence and tobacco smoking per country and sex.
552 Estimate of ASR of tobacco smoking prevalence was obtained from WHO Global Health Observatory
553 (2019). **b**, Association between cigarette quantity smoked per day in the HNC dataset and ASR
554 incidence per country, sex, and subsite. **c**, Association of tobacco-related signatures with ASR
555 incidence. Number of mutations attributed to tobacco-related SBS (SBS4, SBS92, SBS_I), DBS (DBS2,
556 DBS6), and ID (ID3) mutational signatures against ASR of HNC per country, sex, and subsite. For b
557 and c, The p-values shown are for ASR variable in regressions across all cases, adjusted for age. The
558 frequency of the signatures and number of cases per country, sex, and subsite are indicated. OC, oral
559 cavity; OPC, oropharynx.

560
561 **Figure 5. Alcohol-related signatures.** **a**, Mutational burdens of tobacco-related signatures in HNC
562 cases sorted by subsite, alcohol, and tobacco status. Tumor mutational burdens (TMB) per sample is
563 also displayed. For clarity, y axis has been cut at 100,000 for TMB. **b**, Mutational burdens for SBS,
564 DBS, and ID signatures showing positive associations with the tobacco plus alcohol status (n = 265
565 biologically independent samples). The Kruskal–Wallis test (two sided) was used to test for global

566 differences. Box-and-whisker plots are in the style of Tukey. The line within the box is plotted at the
567 median, while upper and lower ends indicate 25th and 75th percentiles. Whiskers show
568 1.5 × interquartile range (IQR), and values outside it are shown as individual data points. Y axes were
569 cut at 1.25 × upper whisker for clarity. Bar plots indicate the frequencies of dichotomized signatures.
570

571 **Figure 6. UV-related signatures in head and neck cancer. a,** Mutational burdens for mutational
572 signatures related to UV light exposure showing positive associations with the HNC anatomical
573 subsite (n = 265 biologically independent samples). The Kruskal–Wallis test (two sided) was used to
574 test for global differences. Box-and-whisker plots are in the style of Tukey. The line within the box is
575 plotted at the median, while upper and lower ends indicate 25th and 75th percentiles. Whiskers show
576 1.5 × interquartile range (IQR), and values outside it are shown as individual data points. Frequencies
577 of positive samples in each category are indicated in bar plots. **b,** Single base substitutions (SBS),
578 doublet base substitutions (DBS) and small insertions and deletions (ID) signature burdens in
579 samples positive for UV exposure based on relative SBS7a-c contributions above 10% of relative
580 mutational burdens. Samples are sorted by lip (inner (n=3) or unspecified (n=1)), tongue, and floor of
581 the mouth location within the oral cavity. Positive tobacco and alcohol status are indicated in black.
582 OC, oral cavity; OPC, oropharynx.

583
584 **Figure 7. Copy number profile and copy number signature analysis in head and neck cancer. a,**
585 Copy number (CN) signatures extracted in 242 HNC tumors. The size of each dot represents the
586 proportion of samples presenting the signature, and the color represents the mean relative
587 attribution of each signature. **b,** Copy number spectrum of the newly-identified signature CN_G,
588 defined by a 48 context copy number classification incorporating loss-of-heterozygosity status, total
589 copy number state, and segment length to categorize segments from allele-specific copy number
590 profiles. **c,** Copy number profiles of HNC cases classified by copy number cluster. Relative signature
591 burdens, CN burden and associated epidemiological characteristics are indicated. The displayed

592 epidemiological variables show significant differences by CN cluster as per Fisher's exact test and
593 Benjamini-Hochberg procedure. **d**, Summary of exposures, driver alterations and CN signatures
594 associated with each CN cluster. Alluvial diagram depicts the frequency of each etiology in the CN
595 clusters. WGD, whole-genome duplication; CIN, chromosomal instability; LOH, loss of
596 heterozygosity.

597 **ONLINE METHODS**

598 **Recruitment of cases and informed consent**

599 The IARC/WHO coordinated participant recruitment through the HEADSpAcE and Central European
600 international networks, comprising 13 collaborators from the eight participating countries in Europe
601 and South America (**Supplementary Table 14**). Inclusion criteria for patients were ≥ 18 years of age
602 (ranging from 18 to 90 years; with a mean of 60 and standard deviation of 12 years), confirmed
603 diagnosis of primary HNC, and no prior cancer treatment. Informed consent was obtained for all
604 participants. Patients were excluded if they had any condition that could interfere with their ability
605 to provide informed consent or if there were no means of obtaining adequate tissues as per protocol
606 requirements. Ethical approvals were first obtained from each local research ethics committee and
607 federal ethics committee when applicable, as well as from the IARC Ethics Committee.

608

609 **Bio-samples and data collection**

610 Dedicated standard operating procedures, following guidelines from the International Cancer
611 Genome Consortium (ICGC), were designed by the IARC/WHO to select adequate retrospective case
612 series with complete biological samples and exposure information as described previously^{1,2}
613 (**Supplementary Table 14**). In brief, for all case series included, anthropometric measures were
614 taken, together with relevant information regarding medical and familial history. All biological
615 samples from retrospective cohorts were collected using rigorous, standardized protocols and
616 fulfilled the required standards of sample collection defined by the IARC/WHO for sequencing and
617 analysis. Retrospective case series were included after examination of their respective recruitment
618 protocols to ensure the availability of necessary biological samples based on standard operating
619 procedures, following guidelines from the ICGC, and also based on the collection of relevant
620 exposure history based on a comparison of validated epidemiological questionnaires from each
621 specific region. Comparable smoking and alcohol history was available from all centers, as well as

622 detailed epidemiological information on oral health, coffee, tea, and mate consumption for specific
623 regions².

624

625 Potential limitations of using retrospective clinical data collected using different protocols from
626 different populations were addressed by central data harmonization to ensure a comparable group
627 of exposure variables (**Supplementary Table 15**). All patient-related data, as well as clinical,
628 demographical, lifestyle, pathological, and outcome data, were pseudonymized locally using a
629 dedicated alphanumeric identifier system before being transferred to the IARC/WHO central
630 database.

631

632 **Expert pathology review**

633 Original diagnostic pathology departments provided diagnostic histological details of contributing
634 cases through standard abstract forms, together with a representative hematoxylin–eosin-stained
635 slide of formalin-fixed paraffin-embedded tumor tissues whenever possible. The IARC/WHO
636 centralized the entire pathology workflow and coordinated a centralized digital pathology
637 examination of frozen tumor tissues collected for the study, as well as formalin-fixed paraffin-
638 embedded sections when available, via a web-based report approach and a dedicated expert panel,
639 following standardized procedures as described previously¹. A minimum of 50% viable tumor cells
640 was required for eligibility for whole-genome sequencing.

641

642 **DNA extraction**

643 Extraction of DNA from fresh frozen tumor and matched blood samples was centrally conducted at
644 IARC/WHO. Of the cases that proceeded to the final analysis ($n=265$), germline DNA was extracted
645 from blood samples using previously described protocols and methods¹.

646

647 **HPV infection status and genome detection**

648 The HPV infection status was determined by HPV16 E1, E2, E6 and E7 serology. To assess the HPV
649 status in oropharynx cases with missing serologic information ($n=3$), we used two orthogonal NGS-
650 based viral integration tools: Virus intEgration sites through iterative Reference SEquence
651 customization (VERSE) and Fast Viral Integration and Fusion Identification (FastViFi)^{3,4}
652 (**Supplementary Table 16; Supplementary Note**). VERSE was utilized as part of the VirusFinder2.0
653 package: <https://bioinfo.uth.edu/VirusFinder/> and FastViFi was installed using github:
654 <https://github.com/sara-javadzadeh/FastViFi>. Default parameters were used for running both tools.

655

656 **Whole-genome sequencing**

657 A total of 618 patients with HNC were enrolled in the study. Out of those, 315 cases were selected
658 based on pathologic review and DNA quality (tumor and germline), and DNA was received at the
659 Wellcome Sanger Institute for whole genome sequencing. To ensure the tumor and matched normal
660 sample originated from the same individual, Fluidigm SNP genotyping with a custom panel was
661 performed. Whole-genome sequencing (150 bp paired-end) was performed on the NovaSeq 6000
662 platform with target coverage of 40× for tumors and 20× for matched normal tissues. All sequencing
663 reads were aligned to the GRCh38 human reference genome using Burrows-Wheeler-MEM (version
664 0.7.16a and version 0.7.17). A standard set of post-sequencing quality criteria was applied for
665 metrics including total coverage, evenness of coverage, and contamination. Cases were excluded if
666 coverage was below 30× for tumors or 15× for normal tissue. For evenness of coverage, the median
667 over mean coverage (MoM) score was calculated, and tumor samples with MoM scores outside the
668 range of values (0.92 – 1.09) which were determined by previous studies to be appropriate were
669 excluded⁵. Conpair⁵ (<https://github.com/nygenome/Conpair>) was used to detect contamination, and
670 any tumor or normal sample with a value above 3% was excluded⁶. A total of 265 cases passed all
671 criteria and were included in subsequent analysis.

672

673 **Somatic variant calling**

674 A standard analysis pipeline (<https://github.com/cancerit>) was used to perform variant calling for
675 copy number variants (ASCAT⁷ and Battenberg⁸, when tumor purity allowed); SNVs (cgpCaVEMan⁹);
676 Indels (cgpPindel¹⁰); and structural rearrangements (BRASS). CaVEMan and BRASS were run using
677 the copy number profile and purity values determined from ASCAT when possible (complete
678 pipeline, $n=242$) or using copy number defaults and an estimate of purity obtained from ASCAT–
679 BATTENBERG when tumor purity was insufficient to determine an accurate copy number profile
680 (partial pipeline, $n=23$). For SNVs, additional filters ($ASRD \geq 140$ and $CLPM = 0$) in addition to the
681 standard PASS filter. To further exclude the possibility of caller-specific artifacts being included in the
682 analysis, a second variant caller, Strelka2, was run for SNVs and IDs^{1,11}, with variants called by both
683 the Sanger variant-calling pipeline and Strelka2 included in the final analysis.

684

685 **Generation of mutational matrices**

686 Mutational matrices for single base substitutions (SBS), doublet base substitutions (DBS), small
687 insertions and deletions (ID), and copy number variants (CNV) were generated using
688 SigProfilerMatrixGenerator (<https://github.com/AlexandrovLab/SigProfilerMatrixGenerator>) with
689 default options (v1.2.0)¹².

690

691 **Mutational signature analysis**

692 Multiple methods were used to extract mutational signatures. The primary extractions were
693 performed using SigProfilerExtractor (<https://github.com/AlexandrovLab/SigProfilerExtractor>) with a
694 second method mSigHdp used to validate the de novo mutational signatures extracted
695 (<https://github.com/steverozen/mSigHdp>)^{13,14}. SigProfilerExtractor v1.1.13 was run using
696 nndsvd_min initialization ($NMF_init="nndsvd_min"$) for 1-20 signature solutions and 500 NMF
697 replicates. For SBS mutational signatures were extracted in both SBS1536 and SBS288 contexts. Both
698 results were similar (**Supplementary Note**) with the SBS1536 results taken forward for the final

699 analysis (**Supplementary Table 4**). Signatures were extracted using SigProfilerExtractor in the
700 following contexts for other variant types; DBS78 for DBS, ID83 for indels, and CNV48 for copy
701 number variants (**Supplementary Tables 5-6,17**). The extracted de novo signatures were
702 decomposed to COSMIC reference signatures where possible; this step is important as it allows the
703 detection of *de novo* signatures which are made up of multiple reference signatures that have not
704 separated during the extraction process (**Supplementary Note**). mSigHdp extractions were
705 performed using the suggested parameters and using the country of origin to construct the
706 hierarchy for SBS96 and ID83 contexts. A comparison of the SigProfilerExtractor and mSigHdp results
707 can be found in the **Supplementary Note**.

708

709 **Attribution of activities of mutational signatures**

710 MSA v2.0 (<https://gitlab.com/s.senkin/MSA>) was used to attribute both de novo and COSMIC
711 mutational signatures¹⁵. For COSMIC attributions the panel of signatures included reference
712 signatures identified during the decomposition of mutational signatures in addition to newly
713 extracted signatures which were not decomposed. A conservative approach was used for MSA
714 attributions utilizing the (params.no_CI_for_penalties=false) option for calculation of optimum
715 penalties. Pruned attributions were used for final analysis, where confidence intervals have been
716 applied to each attributed mutational signature and any signature activity with a lower confidence
717 limit equal to 0 are removed.

718

719 **Driver mutations**

720 Driver mutations in HNC were identified using the following methods. Firstly, dNdS was used to
721 identify genes under positive selection in HNC¹⁶. Results were calculated both for the whole genome
722 (q-value<0.01), and with restricted hypothesis testing (RHT) for a panel of 369 known cancer genes¹⁶.
723 Variants in any gene identified as under positive selection in global dNdS or in the 369-cancer gene
724 panel were considered as potential driver mutations and were then classified as likely drivers if they

725 met any of the following criteria: (i) Truncating mutations in genes annotated as tumor suppressors;
726 (ii) mutations annotated as likely or known oncogenic in MutationMapper; (iii) truncating variants in
727 genes with selection (q value <0.05) for truncating mutations assumed to be tumor suppressors and
728 thus likely drivers; (iv) missense variants in all genes under positive selection and with dN/dS ratios
729 for missense mutations above five (assuming four of every five missense mutations are drivers)
730 labeled as likely drivers; or (v) in-frame indels in genes under significant positive selection for in-
731 frame indels. The Cancer Gene Census (<https://cancer.sanger.ac.uk/census>) and the Cancer Genome
732 Interpreter tool (<https://www.cancergenomeinterpreter.org>) were used to annotate potential
733 drivers with the mode of action. Missense mutations were assessed using the MutationMapper tool
734 (http://www.cbioportal.org/mutation_mapper).

735

736 **Copy number profile**

737 The copy number profiles were investigated in a subset of cases with available copy number data
738 (complete pipeline, $n=242$). Unsupervised clustering analysis of the copy number counts was
739 performed using Euclidean distance and Ward's agglomerative procedure. Driver copy number
740 alterations were defined as cancer-related alterations in the COSMIC cancer gene census as
741 follows^{17,18}: (1) homozygous deletion (CN = (0, 0)) of genes listed as deleted in COSMIC; and (2)
742 amplification (CN $> 2 \times$ ploidy + 1) of genes listed as amplified (A) in COSMIC or *PIK3CA* gains, a
743 commonly-reported HNC alteration^{19,20}.

744

745 **Evolutionary analysis**

746 MutationTimer²¹ was run to annotate mutations as either early clonal, late clonal, subclonal, or NA
747 clonal (meaning clonality could not be assigned). Samples with at least 256 early clonal mutations
748 and at least 256 late clonal mutations were retained ($n=173$), and the early and late clonal mutations
749 for these samples were split into individual VCF files. SigProfilerAssignment²² was run on the
750 resulting VCF files to identify the mutational processes active in the early clonal and late clonal

751 mutations for each sample. Differences between the early and late relative activity of each
752 mutational signature were assessed using a Wilcoxon signed-rank test, and p-values were corrected
753 across signatures using the Benjamini-Hochberg Procedure (q-value).

754

755 **Regressions and associations with signatures**

756 Signature attributions were dichotomized into presence and absence using confidence intervals,
757 with presence defined as both lower and upper limits being positive, and absence as the lower limit
758 being zero. If a signature was present in at least 75% of cases (SBS1, SBS2, SBS13, SBS18, SBS_I, ID1,
759 and ID2), it was dichotomized into above and below the median of attributed mutation counts. The
760 binary attributions served as dependent variables in logistic regressions, and relevant risk factors
761 were used as factorized independent variables. Regressions with variables presenting data
762 separation were performed using Firth's penalized logistic regression.

763

764 For SBS, DBS, and ID mutation burden analyses, cases defined as hypermutators (mutation burdens
765 more than 1.5 IQR above Q3) were excluded and associations with epidemiological factors were
766 assessed using linear regression analysis.

767

768 Regressions with HNC incidence (age-standardized rates) were performed as linear regressions with
769 signature attributions (those present in at least 75% of cases) with confidence intervals not
770 consistent with zero. Signatures present in less than 75% of cases were dichotomized into presence
771 and absence as previously mentioned and analyzed using the logistic regressions. Age-standardized
772 rates were obtained from Global Cancer Observatory (GLOBOCAN 2022)²³. Regressions were
773 performed on a sample basis.

774

775 To adjust for confounding factors, sex, age of diagnosis, subsite, region, tobacco, and alcohol status
776 were added as covariates in all regressions. The region variable was categorized as Europe and South
777 America. The Bonferroni method was used to test for significant p-values.

778

779 **DATA AVAILABILITY**

780 Whole genome sequencing data and patient metadata are deposited in the European Genome-
781 phenome Archive (EGA) associated with study EGAS00001005450. Mutational catalogs for the
782 PCAWG dataset can be accessed at <https://dcc.icgc.org/releases/PCAWG>. All other data is provided
783 in the accompanying Supplementary Tables.

784

785 **CODE AVAILABILITY**

786 All algorithms used for data analysis are publicly available with repositories noted within the
787 respective method sections and in the accompanying reporting summary. Code used for regression
788 analysis and figures is available at <https://gitlab.com/mutographs-hnc>.

789 **METHODS REFERENCES**

- 790 1. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight
791 countries with varying incidence. *Nat Genet* **53**, 1553–1563 (2021).
- 792 2. Perdomo, S. *et al.* The Mutographs biorepository: A unique genomic resource to study
793 cancer around the world. *Cell Genomics* 100500 (2024).
- 794 3. Javadzadeh, S. *et al.* FastViFi: Fast and accurate detection of (Hybrid) Viral DNA and RNA.
795 *NAR Genom Bioinform* **4**, (2022).
- 796 4. Wang, Q., Jia, P. & Zhao, Z. VERSE: A novel approach to detect virus integration in host
797 genomes through reference genome customization. *Genome Med* **7**, 1–9 (2015).
- 798 5. Whalley, J. P. *et al.* Framework for quality assessment of whole genome cancer sequences.
799 *Nat Commun* **11**, (2020).
- 800 6. Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and
801 contamination estimator for matched tumor–normal pairs. *Bioinformatics* **32**, 3196 (2016).
- 802 7. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*
803 **107**, 16910–16915 (2010).
- 804 8. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
- 805 9. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect
806 Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15.10.1-
807 15.10.18 (2016).
- 808 10. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events
809 from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15.7.1 (2015).
- 810 11. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature*
811 *Methods* 2018 15:8 **15**, 591–594 (2018).
- 812 12. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring
813 patterns of small mutational events. *BMC Genomics* **20**, (2019).

- 814 13. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with
815 SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
- 816 14. Liu, M., Wu, Y., Jiang, N., Boot, A. & Rozen, S. G. mSigHdp: hierarchical Dirichlet process
817 mixture modeling for mutational signature discovery. *NAR Genom Bioinform* **5**, (2023).
- 818 15. Senkin, S. MSA: reproducible mutational signature attribution with confidence based on
819 simulations. *BMC Bioinformatics* **22**, 1–11 (2021).
- 820 16. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*
821 **171**, 1029-1041.e21 (2017).
- 822 17. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **2022**
823 *606:7916* **606**, 984–991 (2022).
- 824 18. Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*
825 **47**, D941–D947 (2019).
- 826 19. Sayáns, M. P. *et al.* Comprehensive Genomic Review of TCGA Head and Neck Squamous Cell
827 Carcinomas (HNSCC). *Journal of Clinical Medicine* **2019**, Vol. 8, Page 1896 **8**, 1896 (2019).
- 828 20. Leemans, C. R., Snijders, P. J. F. & Brakenhoff, R. H. The molecular landscape of head and
829 neck cancer. *Nat Rev Cancer* **18**, 269–282 (2018).
- 830 21. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **2020** *578:7793* **578**,
831 122–128 (2020).
- 832 22. Díaz-Gay, M. *et al.* Assigning mutational signatures to individual samples and individual
833 somatic mutations with SigProfilerAssignment. *bioRxiv* (2023)
834 doi:10.1101/2023.07.10.548264.
- 835 23. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality
836 worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **71**, 209–249 (2021).
- 837
- 838
- 839

840 **ACKNOWLEDGEMENTS**

841 The authors would like to thank Laura O'Neill, Kirsty Roberts, Katie Smith, Siobhan Austin-Guest and
842 the staff of Sequencing Operations at the Wellcome Sanger Institute for their contribution. We are
843 grateful for the support provided by the IARC General Services, including the Laboratory Services and
844 Biobank team led by Z. Kozlakidis, the Section of Support to Research overseen by T. Landes. The
845 authors would like to thank Maggie Blanks and Mimi McCord for useful discussions. The authors
846 would also like to thank all the patients and their families involved in this study.

847

848 **FUNDING**

849 This work was delivered as part of the Mutographs team supported by the Cancer Grand Challenges
850 partnership funded by Cancer Research UK (C98/A24032). Work at the Wellcome Sanger Institute
851 was also supported by the Wellcome Trust (grants 206194 and 220540/Z/20/A), and work at the
852 IARC/WHO was supported by regular budget funding. This work was also supported by the US
853 National Institute of Health grants R01ES032547-01, R01CA269919-01, and 1U01CA290479-01 to
854 L.B.A. as well as by L.B.A.'s Packard Fellowship for Science and Engineering. The research performed
855 in L.B.A.'s lab was supported by UC San Diego Sanford Stem Cell Institute. The head and neck cancer
856 collection received funding from the European Union's Horizon 2020 research and innovation
857 program under grant no. 825771 and the São Paulo Research Foundation, FAPESP 2018/26297-3. J.P.
858 and F.G. were partially supported by Italian Ministry of Health - Ricerca Corrente. The funders had
859 no roles in study design, data collection and analysis, decision to publish, or preparation of the
860 manuscript.

861

862 **CONTRIBUTIONS**

863 The study was conceived and designed by S.P. and P.B., and supervised by S.P., P.B., M.R.S., and
864 L.B.A. Analysis of data was performed by L.T., S.M., A.C.D.C., M.K., S.C., S.S., T.C., R.C.C.P., J.R.A.,
865 V.G., A.A., E.N.B., R.V., J.W., S.F., M.D., D.J., and J.T. Pathology review was carried out by B.A. Sample

866 manipulation was carried out by P.C., C.Carreira, and C.L. Patient and sample recruitment was led or
867 facilitated by F.R.P., L.P.K., J.P., F.G., J.C.D.C., P.L., A.L., M.V., D.M., I.N.M., L.M.A., R.R., J.R.V.P.,
868 S.V.V.Z., I.H., M.P.C., C.Canova, E.F., and P.A.R. Data generation was performed by J.W., S.F., and C.L.
869 Scientific project management was carried out by A.C.D.C. and L.H. L.T. L.T. and S.M. contributed and
870 were responsible for overall scientific coordination. The manuscript was written by L.T., S.M.,
871 A.C.D.C., L.B.A., P.B., M.R.S., and S.P., with contributions from all other authors.

872

873 **COMPETING INTERESTS**

874 L.B.A. is a co-founder, CSO, scientific advisory member, and consultant for io9, has equity and
875 receives income. The terms of this arrangement have been reviewed and approved by the University
876 of California, San Diego in accordance with its conflict of interest policies. L.B.A. is also a
877 compensated member of the scientific advisory board of Inocras. L.B.A.'s spouse is an employee of
878 Biotheranostics. E.N.B. and L.B.A. declare U.S. provisional patent applications filed with UCSD with
879 serial numbers 63/289,601; 63/269,033; and 63/483,237. A.A. and L.B.A. declare U.S. provisional
880 patent application filed with UCSD with serial number 63/366,392. L.B.A. also declares U.S.
881 provisional application 63/412,835 and international application PCT/US2023/010679 filed with
882 UCSD. L.B.A. is also an inventor of a US Patent 10,776,718 for source identification by non-negative
883 matrix factorization. All other authors declare that they have no competing interests.

884

885 **DISCLAIMER**

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

890

891 **CORRESPONDING AUTHOR**

892 Correspondence to Sandra Perdomo.

893 **EXTENDED DATA FIGURE LEGENDS**

894 **Extended Data Figure 1. Mutational burdens in HNC. a-c,** Mutational burdens for single base
895 substitutions (SBS), doublet base substitutions (DBS) and small insertions and deletions (ID) burdens
896 by anatomical subsite (a), smoking status (b) and country (c). Panel b depicts the mutation burdens
897 by smoking status in the whole HNC dataset (left) and across anatomical subsites (right). Kruskal–
898 Wallis test (two sided) was used to test for global differences. Box-and-whisker plots are in the style
899 of Tukey. The line within the box is plotted at the median, while upper and lower ends indicate 25th
900 and 75th percentiles. Whiskers show 1.5 × interquartile range (IQR), and values outside it are shown
901 as individual data points. Hypermutators defined as samples with mutation burdens above 100,000
902 for SBS ($n=4$), 6,000 for DBS ($n=1$) and 5,000 for ID ($n=1$) were removed from the analysis. OC, oral
903 cavity; OPC, oropharynx; HPX, hypopharynx.

904
905 **Extended Data Figure 2. SBS signature decomposition.** Decomposed SBS signatures, including
906 reference COSMIC signatures and *de novo* signatures not decomposed into COSMIC reference
907 signatures.

908
909 **Extended Data Figure 3. DBS and ID signature decomposition.** Decomposed DBS (a) and ID (b)
910 signatures, including reference COSMIC signatures and *de novo* signatures not decomposed into
911 COSMIC reference signatures.

912
913 **Extended data Figure 4. Evolutionary analysis of mutational signatures and driver mutations in**
914 **HNC. a,** Comparison of mutational signatures between early and late clonal mutations in HNC
915 ($n=173$). **b,** Relative activities of SBS_I in early and late clonal mutations across anatomical subsites.
916 Lines show the change in relative activity between the early and late clonal mutations within a
917 positive sample. Colored lines represent an activity change of more than 6% (blue indicates higher in
918 the clonal early mutations; orange indicates higher in the clonal late mutations). Bar plots show the

919 distribution of activities in samples where the signature was present in the early and/or late clonal
920 mutations; the number of positive samples is represented in the title of each plot. Black bars
921 indicate one standard deviation away from the mean. Significance was assessed using a two-sided
922 Wilcoxon signed-rank test, and p-values were corrected using the Benjamini-Hochberg Procedure (q-
923 value). OC, oral cavity; OPC, oropharynx; LYX, larynx; HPX, hypopharynx.

924

925 **Extended Data Figure 5. Association of tobacco-related mutational signatures with anatomical**

926 **subsites and tobacco status. a,** Mutational burdens for tobacco-related mutational signatures by

927 anatomical subsite ($n = 265$ biologically independent samples). **b,** Mutational burdens for mutational

928 signatures showing significant negative associations with tobacco consumption. The Kruskal–Wallis

929 test (two sided) was used to test for global differences. Box-and-whisker plots are in the style of

930 Tukey. The line within the box is plotted at the median, while upper and lower ends indicate 25th and

931 75th percentiles. Whiskers show $1.5 \times$ interquartile range (IQR), and values outside it are shown as

932 individual data points. Y axes were cut at $1.25 \times$ upper whisker for clarity. Bar plots indicate the

933 frequencies of dichotomized signatures OC, oral cavity; OPC, oropharynx.

934

935 **Extended Data Figure 6. Driver alterations and driver mutation spectra in HNC. a,** Driver mutations

936 in HNC samples ($n=265$) sorted by tobacco and alcohol status. Genes mutated in more than 2% of

937 the cases are shown. **b,** Driver mutations and copy number events in HNC samples with available

938 copy number data ($n=242$). Only driver genes with both copy number gains and losses are included.

939 Top, tumor mutational burden (TMB) per sample. Middle, presence of mutations per sample.

940 Bottom, epidemiological characteristics. Frequency of mutations in the HNC dataset and q values

941 from two-sided Fisher's exact test are displayed. **c,** SBS96-mutation spectrum of driver mutations in

942 smokers and non-smoker HNC cases and percentage of driver mutations occurring in C>A contexts.

943

944 **Extended Data Figure 7. Mutational signature and driver spectra of oropharynx HNC cases by HPV**
945 **status. a**, Average relative attributions of SBS signatures by human papillomavirus (HPV) positivity
946 and tobacco status in oropharynx (OPC) cancers. **b**, Driver mutations in OPC HNC samples ($n=46$)
947 sorted by HPV and tobacco status. Genes mutated in more than 2% of the samples are shown. **c**,
948 Driver mutations and copy number events in OPC HNC samples with available copy number data
949 ($n=44$). Only driver genes with copy number gains and losses are included. Top, tumor mutational
950 burden (TMB) per sample. Middle, presence of mutations per sample. Bottom, HPV status and
951 tobacco smoking. Frequency of mutations in the HNC dataset and q values from two-sided Fisher's
952 exact test are displayed.

953
954 **Extended Data Figure 8. Copy number profile of head and neck cancer clusters. a**, Genome-wide
955 segments showing major and minor allele counts in 10 randomly picked samples per copy number
956 (CN) cluster. **b**, Ploidy, CN burden, and burden of CN gains, losses, and CN neutral LOH (NLOH) across
957 clusters ($n = 242$ biologically independent samples). The Kruskal–Wallis test (two sided) was used to
958 test for global differences. Box-and-whisker plots are in the style of Tukey. The line within the box is
959 plotted at the median, while upper and lower ends indicate 25th and 75th percentiles. Whiskers show
960 $1.5 \times$ interquartile range (IQR), and values outside it are shown as individual data points.

961
962 **Extended Data Figure 9. Copy number signature decomposition.** Decomposed copy number
963 signatures including reference COSMIC signatures and *de novo* signatures not decomposed into
964 COSMIC reference signatures.

965
966 **Extended Data Figure 10. Copy number signature enrichment by HNC clusters and driver profile. a-**
967 **b**, Signature burdens for copy number signatures by copy number cluster showing associations with
968 cluster D (**a**) and cluster P (**b**). **c**, Signature burdens for CN5 and CN_G signatures in copy number
969 clusters D and P. The Kruskal–Wallis test (two sided) was used to test for global differences. Box-and-

970 whisker plots are in the style of Tukey. The line within the box is plotted at the median, while upper
971 and lower ends indicate 25th and 75th percentiles. Whiskers show $1.5 \times$ interquartile range (IQR), and
972 values outside it are shown as individual data points. Y axis were cut at $1.25 \times$ upper whisker for
973 clarity. Bar plots indicate the frequencies of dichotomized signatures. **d**, Associations between copy
974 number clusters or signatures and driver alterations. Effect size ($\log_2(\text{OR})$, color), and significance
975 level ($-\log_2(q)$, size) from two-sided Fisher's exact tests, corrected using the Benjamini-Hochberg
976 Procedure, are displayed. Only significant associations are shown ($q < 0.05$).

977 **SUPPLEMENTARY TABLE AND FIGURE LEGENDS**

978 **Supplementary Table 1. Clinical and epidemiological characteristics of the HNC dataset.**

979 **Supplementary Table 2. Mutation burdens per sample.**

980 **Supplementary Table 3. Associations of HNC risk factors with and mutational burden.**

981 **Supplementary Table 4. SBS *de novo* signatures extracted from 265 HNC cases.**

982 **Supplementary Table 5. DBS *de novo* signatures extracted from 265 HNC cases.**

983 **Supplementary Table 6. ID *de novo* signatures extracted from 265 HNC cases.**

984 **Supplementary Table 7. Decomposition of *de novo* mutational signatures to COSMIC reference**
985 **signatures.**

986 **Supplementary Table 8. Activities of *de novo* mutational signatures in HNC cases.**

987 **Supplementary Table 9. Activities of decomposed COSMIC signatures in HNC cases.**

988 **Supplementary Table 10. Associations of HNC risk factors with COSMIC mutational signatures.**

989 **Supplementary Table 11. Likely driver mutations identified in HNC.**

990 **Supplementary Table 12. Likely driver copy number alterations identified in HNC.**

991 **Supplementary Table 13. Associations of HNC risk factors with copy number clusters.**

992 **Supplementary Table 14. Details of individual case collections.**

993 **Supplementary Table 15. Details of data harmonization for HNC risk factors.**

994 **Supplementary Table 16. Details of HPV16 assessment in oropharynx cases.**

995 **Supplementary Table 17. CNV48 *de novo* signatures extracted from 242 HNC cases.**

996

997 **Supplementary Figure 1. Correlations amongst mutational signatures.** Pearson correlation
998 coefficients for each significant comparison are indicated.

999

1000 **Supplementary Figure 2. Association of tobacco-related signatures and HNC incidence.** Association
1001 between tobacco-related signatures with age-standardized rate (ASR) incidence. Number of
1002 mutations attributed to SBS4, SBS92, SBS_I, DBS2, DBS6, and ID3 mutational signatures against ASR

1003 of HNC per country, sex, and subsite. The p-values shown are for ASR variable in regressions across
1004 all cases, adjusted for age. The frequency of the signatures and number of cases per group are
1005 indicated.

1006

1007 **Supplementary Figure 3. Hierarchical clustering of copy number data.** Unsupervised hierarchical
1008 clustering analysis of copy number counts in the HNC cohort ($n=242$) using Euclidean distance and
1009 Ward's agglomerative procedure. Two main clusters (diploid (D) and polyploid (P)) were obtained,
1010 which further subdivided into four groups. Right panel shows the copy number frequency in the HNC
1011 cohort.

1012

1013 **Supplementary Figure 4. Single base substitution signatures extracted by SigProfilerExtractor.** All
1014 single base substitution (SBS) de novo signatures extracted in SBS-1536 (15 signatures) and SBS-288
1015 (14 signatures) format, shown side by side for comparison. Equivalent signatures were not
1016 extracted in SBS-288 format for SBS1536J. For clarity, the signatures context is retained in the
1017 signature names in this figure.

1018

1019 **Supplementary Figure 5. Single base substitution mutational signatures extracted by mSigHdp.**
1020 Fifteen single bases substitution (SBS) de novo signatures extracted by mSigHdp.

1021

1022 **Supplementary Figure 6. Small insertion and deletion mutational signatures extracted by mSigHdp.**
1023 Eight small insertion and deletion (ID) de novo signatures extracted by mSigHdp.

1024

1025 **Supplementary Figure 7. Mutational spectra supporting non-decomposed mutational signatures.**
1026 Individual mutational spectra are shown for cases which support the existence of non-decomposed
1027 signatures SBS_I (SBS1536I) (a), SBS_L (SBS1536_L) (b) and DBS_D (DBS78D) (c).

1028

1029 **Supplementary Figure 8. Principal component analysis of HNC SBS96 mutation counts and**
1030 **signature attributions.** Principal component analysis (PCA) performed on 256 cases of HNC on
1031 relative SBS96 mutation counts colored by **a**, anatomic site, **b**, tobacco status, and **c**, relative
1032 proportion of each mutation class (C>A, C>G, C>T, T>A, T>C, T>G). Circled on the anatomic site/ C>T
1033 plot is a subset of oral cavity HNC which have UV exposure. PCA performed on 256 cases of HNC on
1034 relative signature attributions colored by **d**, anatomic site, **e**, tobacco status and **f**, relative
1035 attributions of tobacco associated signatures SBS4, SBS92 and SBS_I.

1036

1037 **Supplementary Figure 9. UV exposure in HNC.** Support for the presence of UV in HNC of the oral
1038 cavity showing **a**, representative HNC oral cavity mutational specter which is consistent with
1039 representative melanoma mutational spectra from the PCAWG cohort and **b**, correlation between
1040 mutational signatures known to be associated with UV exposure in HNC. Correlation coefficients for
1041 each significant comparison are indicated.

1042

1043 **Supplementary Figure 10. Attribution of HNC mutational signatures in external datasets.**
1044 Attribution of HNC mutational signatures SBS_I (**a**) and SBS_L (**b**) in external datasets. The Kruskal–
1045 Wallis test (two sided) was used to test for global differences. Box-and-whisker plots are in the style
1046 of Tukey. The line within the box is plotted at the median, while upper and lower ends indicate 25th
1047 and 75th percentiles. Whiskers show 1.5 × interquartile range (IQR), and values outside it are shown
1048 as individual data points. Overall mutational signature landscape in in the external datasets was
1049 similar (**c**) with the presence of additional individual mutational spectra (**d**) supporting the existence
1050 of SBS_I.

1051

1052 **Supplementary Figure 11. Correlations amongst copy number signatures.** Pearson correlation
1053 coefficients for each significant comparison are indicated.

1054

1055 **Supplementary Figure 12. Predicted ancestry in HNC.** **a**, Scatter plots of principal components PC1
1056 and PC2 based on genotype data showing the genetic structure of the HNC cohort across different
1057 countries of origin. **b**, Ancestry admixture in the HNC cohort. **c**, Probability of African ancestry by
1058 country.

1059
1060 **Supplementary Figure 13. Clustered mutations in HNC.** a-b, Distribution of clustered mutations in
1061 HNC by tobacco status (a) and anatomical subsite (b) ordered by median tumor mutational burden
1062 (TMB). Each dot represents a single tumor. The clustered mutation ratio is calculated as the fraction
1063 of clustered mutations compared to the total number of mutations in a given sample. Each clustered
1064 event is subclassified and summarized as the proportion of mutations per country associated with a
1065 double-base substitution event, an omikli event, or as a kataegis event.

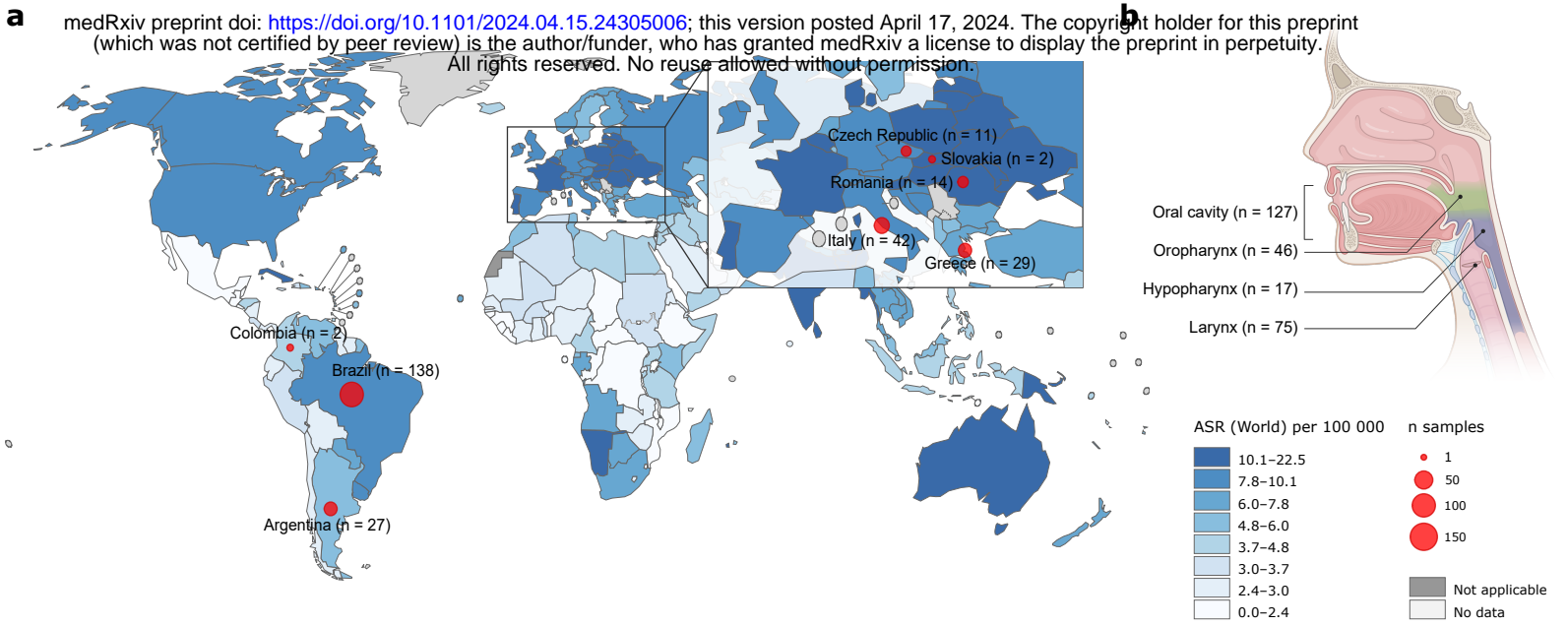
1066
1067 **Supplementary Figure 14. Evolutionary analysis of driver mutations in HNC.** Relative frequency of
1068 driver mutations across early clonal and late clonal/subclonal stages, for the most common driver
1069 genes in HNC ($n=173$).

1070
1071 **Supplementary Figure 15. Human papillomavirus integration in HNC tumors.** **a**, Frequency of
1072 HPV16 integrations in genic and non-genic regions. **b**, Integration sites detected in chromosomal,
1073 cytoband, and genic regions. Rows represent samples positive for viral integration. The number of
1074 integrations per site and sample is depicted. Only four samples presented integrations in genic
1075 regions. **c-d**, Circos plots representing viral integration sites, structural variations (SV) and copy
1076 number (CN) alterations in tumor genomes presenting HPV16 integration. HPV integrations (in
1077 yellow) are depicted in the outermost ring, CN in the inner ring, and SV events in the center. Specific
1078 SV and CN alterations (CNA) surrounding the sites of integration (dotted lines) are shown for three
1079 samples (**d**).

1080

1081 **Supplementary Figure 16. Associations between germline *ADH1B* and *ADH7* variant genotype and**
1082 **alcohol related mutational signatures. *ADH1B* rs1229984 and *ADH7* rs1573496 germline variant**
1083 **genotypes for signatures SBS16 (a), DBS4 (b) and ID11 (c) ($n=265$ biologically independent samples).**
1084 Mutated samples correspond to those with at least one alternative allele. The Kruskal–Wallis test
1085 (two sided) was used to test for global differences. Box-and-whisker plots are in the style of Tukey.
1086 The line within the box is plotted at the median, while upper and lower ends indicate 25th and 75th
1087 percentiles. Whiskers show $1.5 \times$ interquartile range (IQR), and values outside it are shown as
1088 individual data points. Y axes were cut at $1.25 \times$ upper whisker for clarity. Bar plots indicate the
1089 frequencies of dichotomized signatures.

Figure 1



All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

Cancer TODAY | IARC
<https://gco.iarc.who.int/today>
Data version: Globocan 2022 - 08.02.2024
© All Rights Reserved 2024

International Agency
for Research on Cancer
World Health
Organization

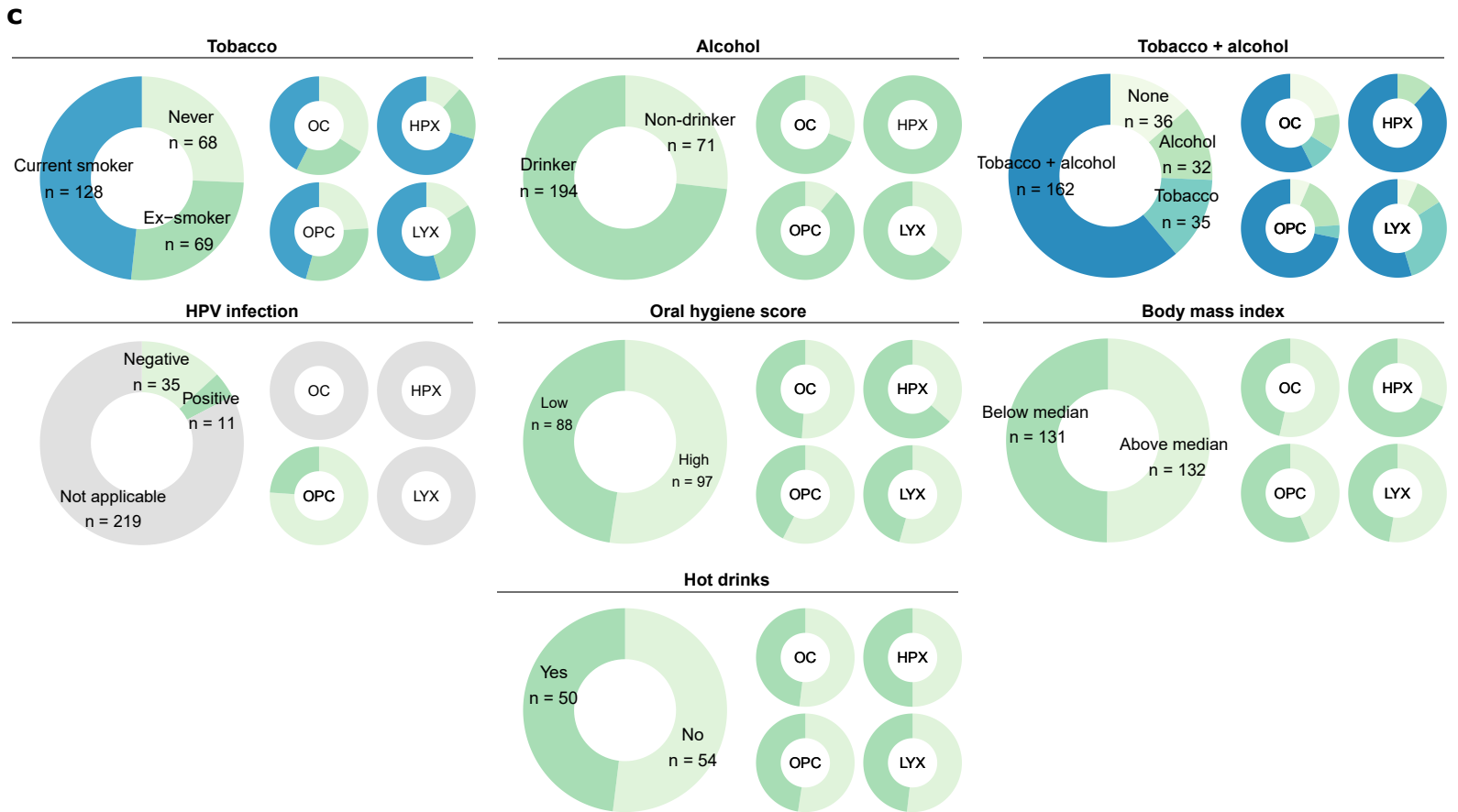


Figure 2

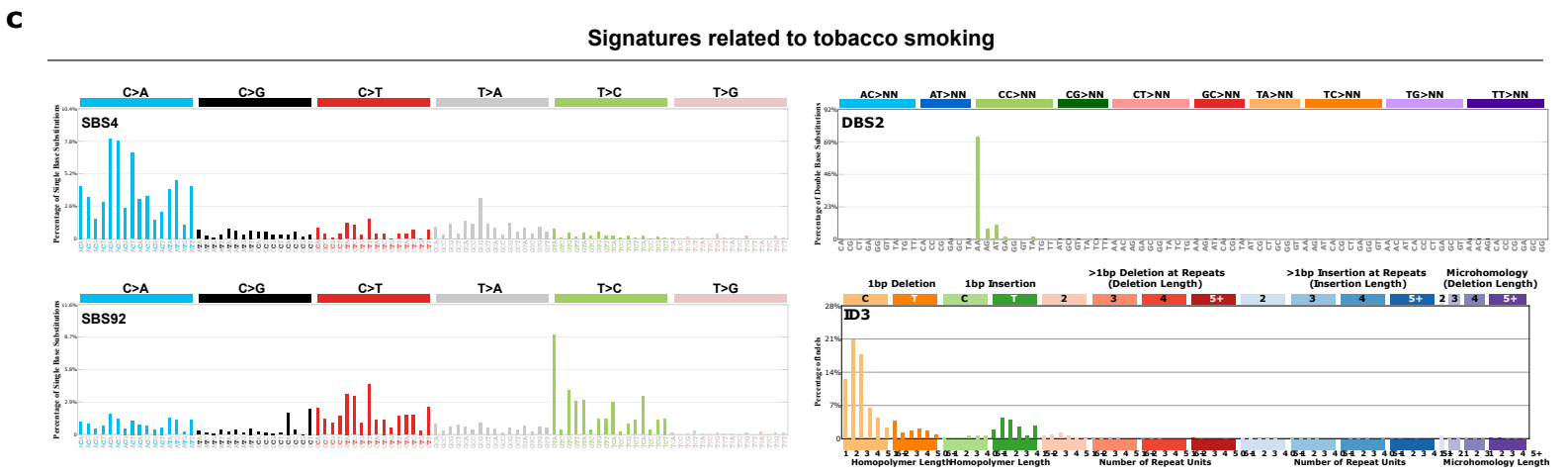
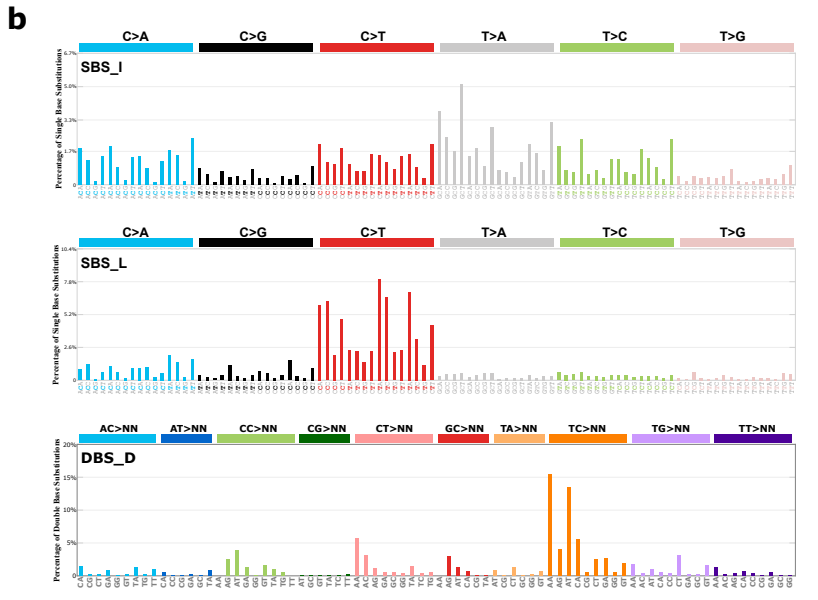
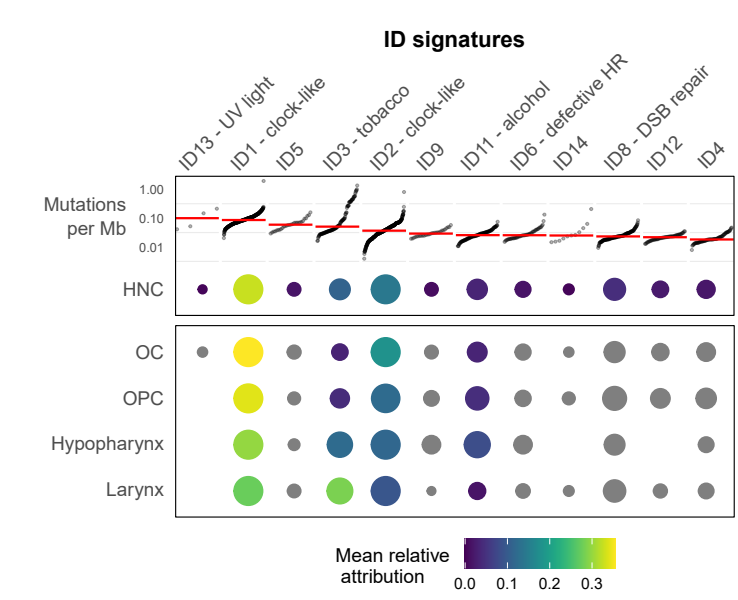
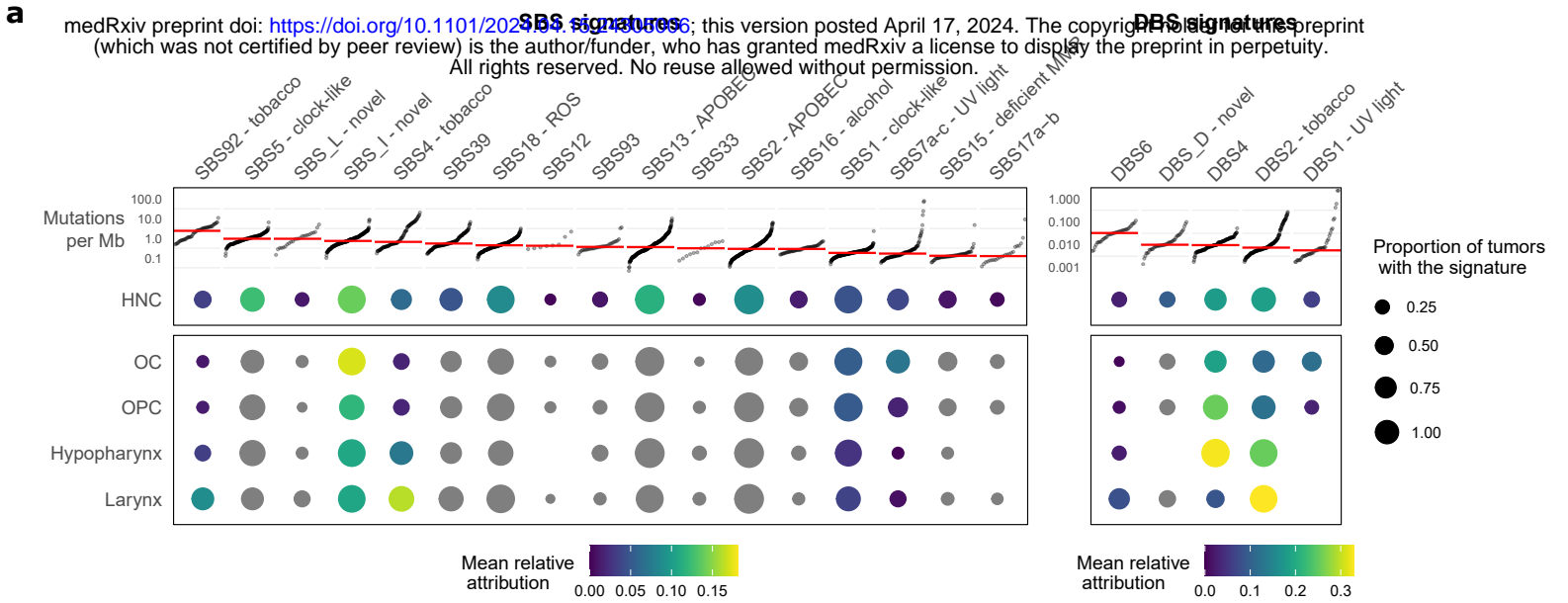


Figure 3

a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

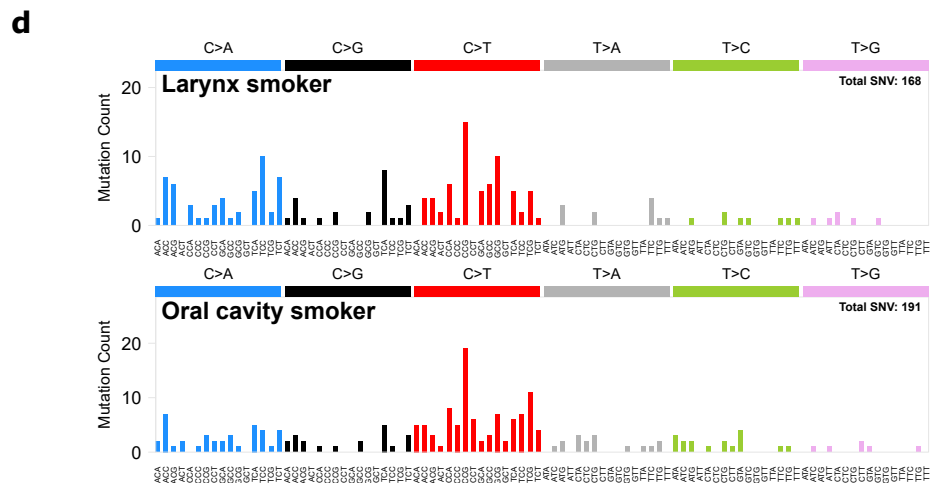
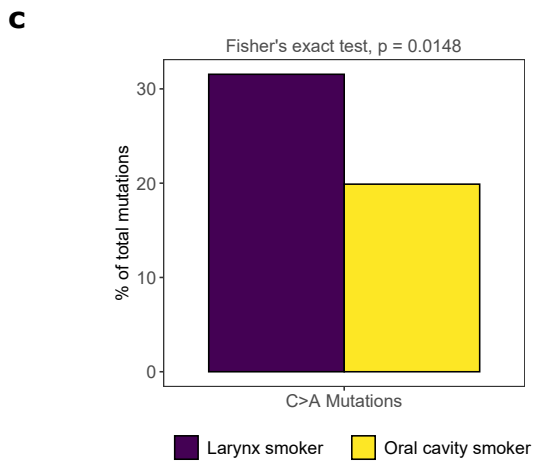
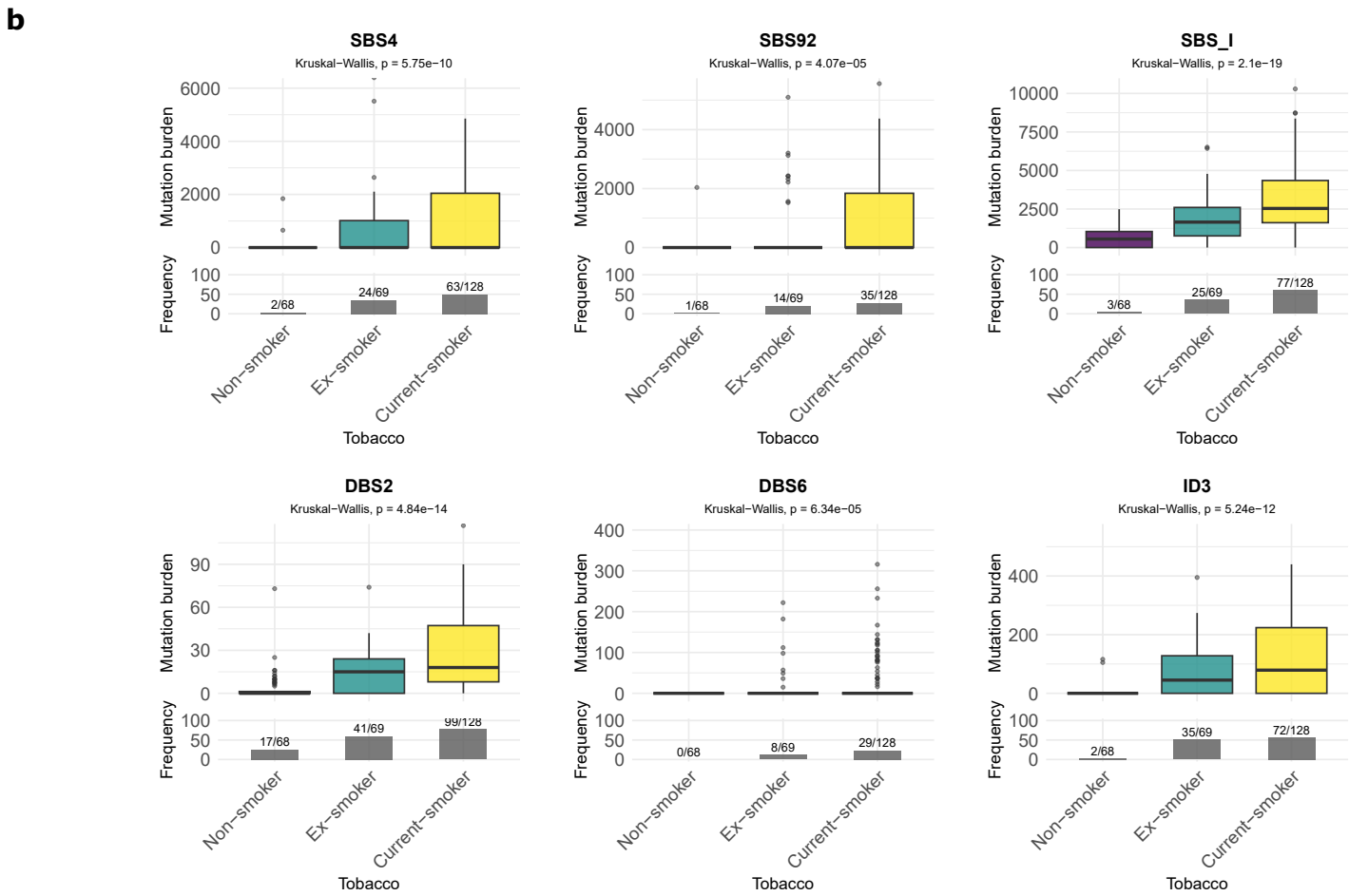
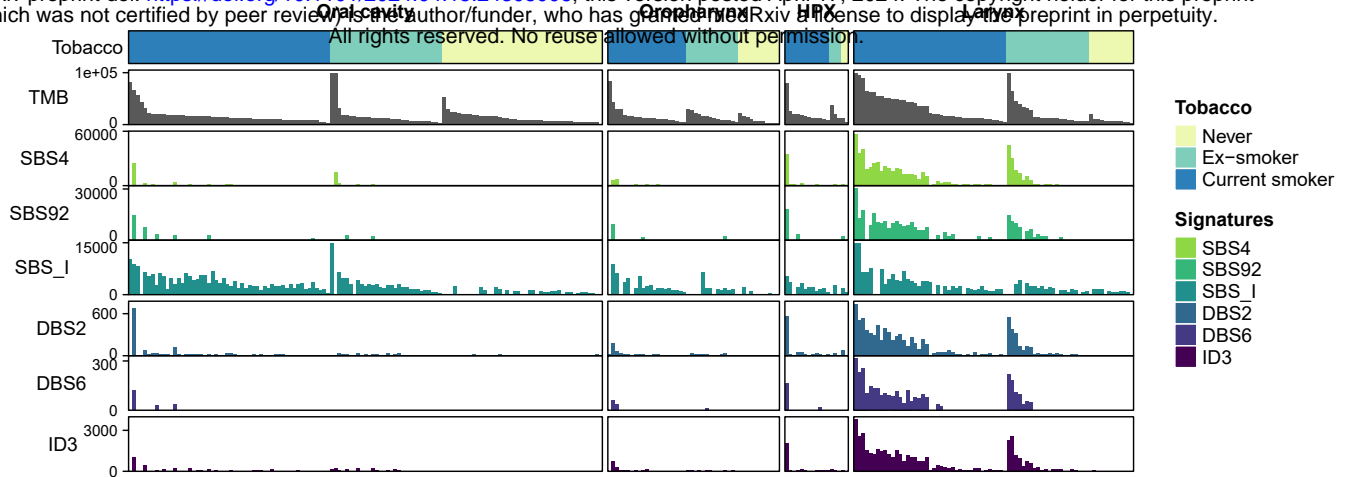


Figure 4

a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305000>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

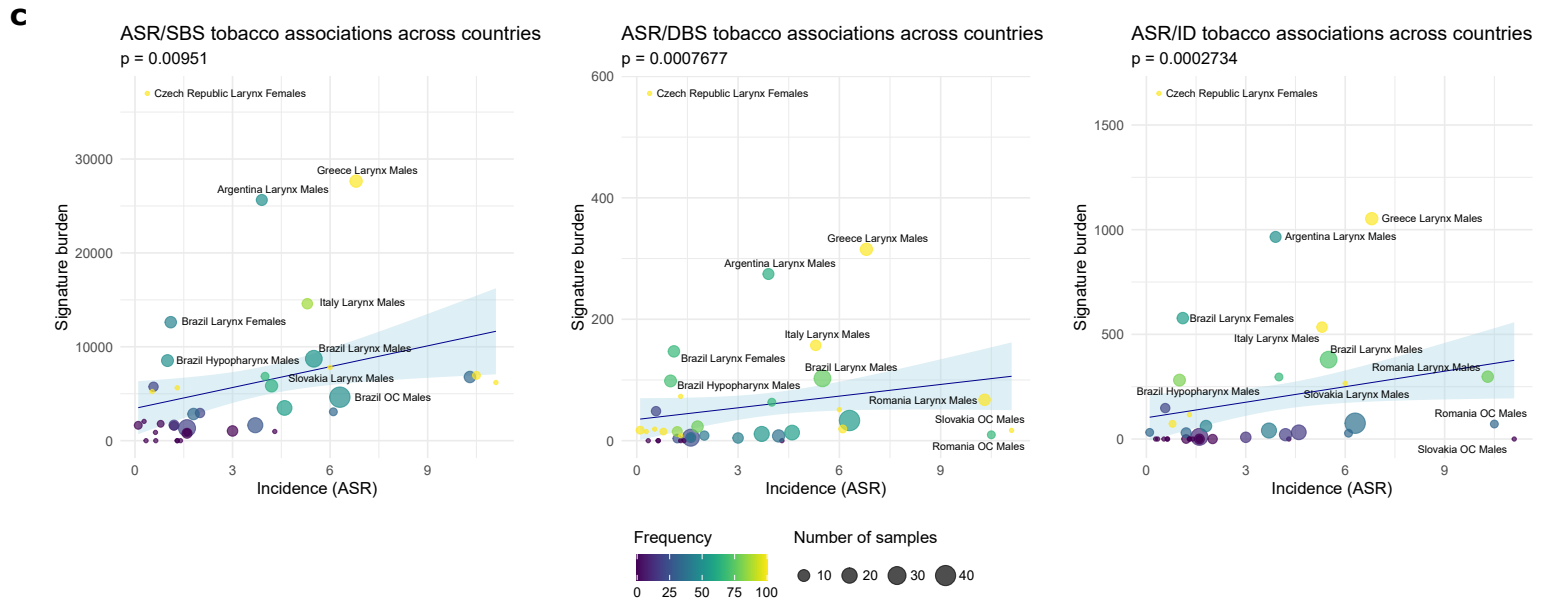
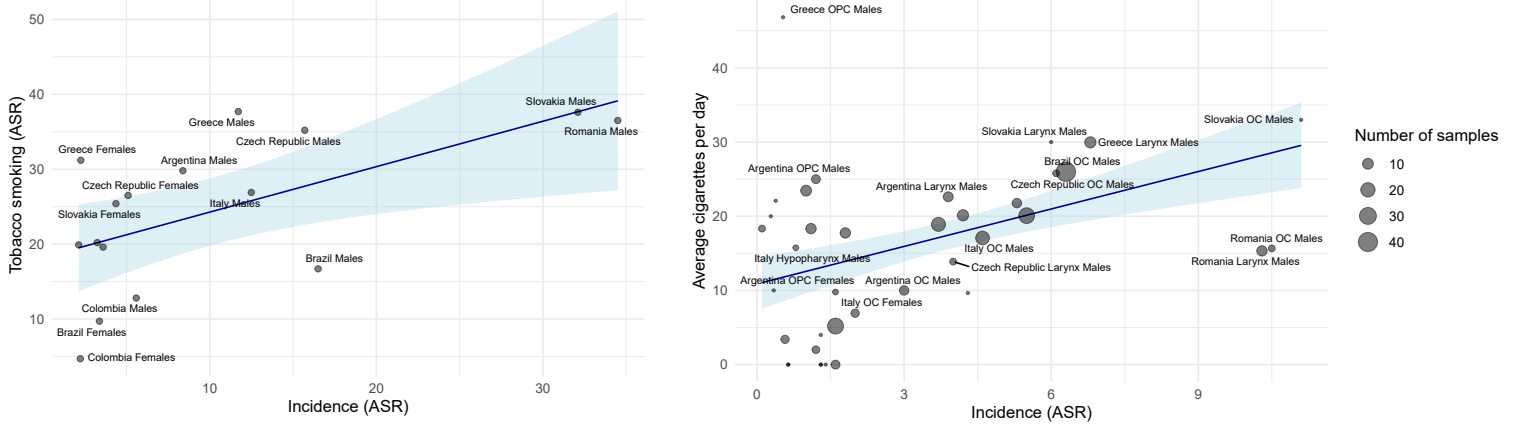


Figure 5

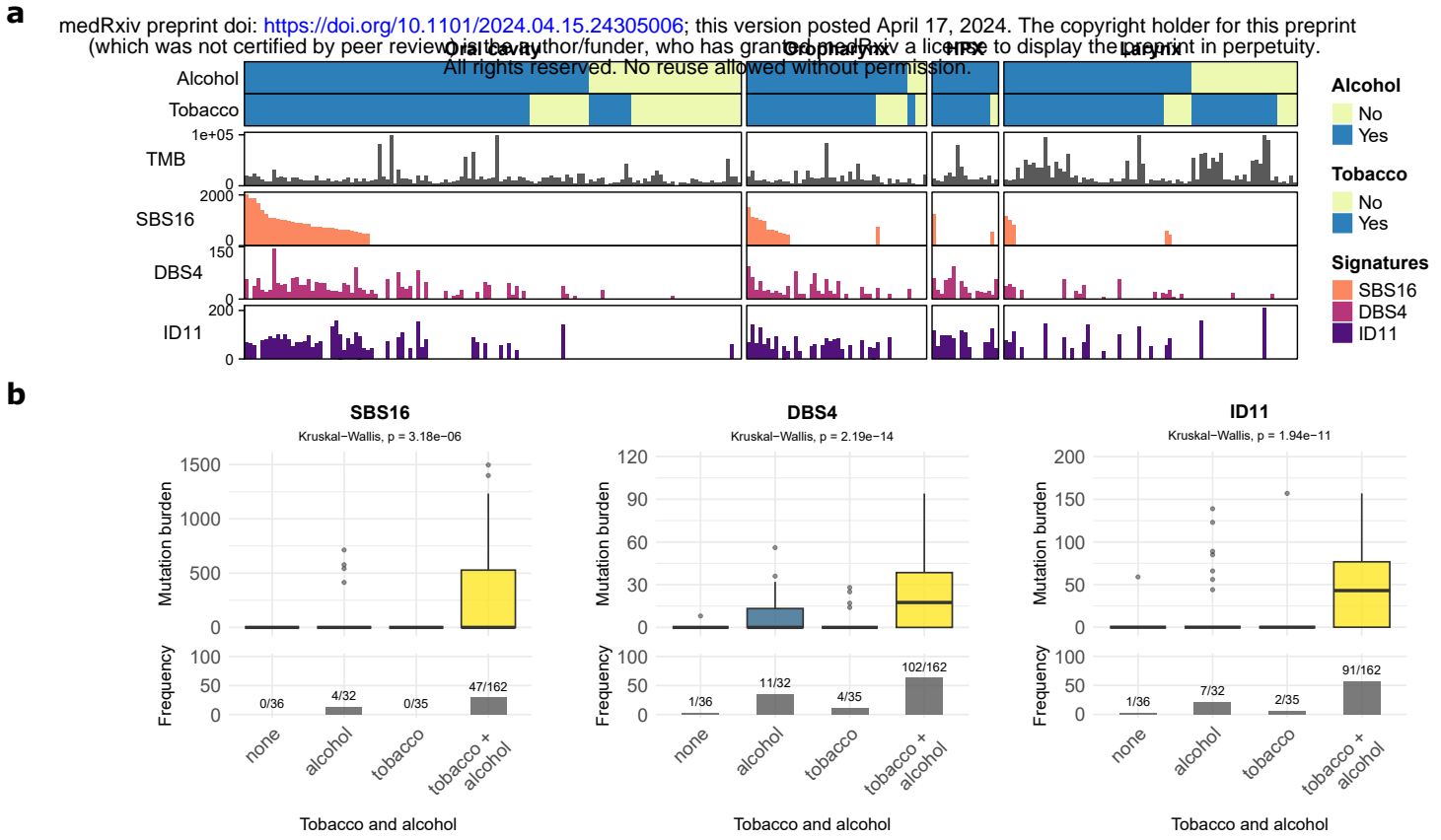


Figure 6

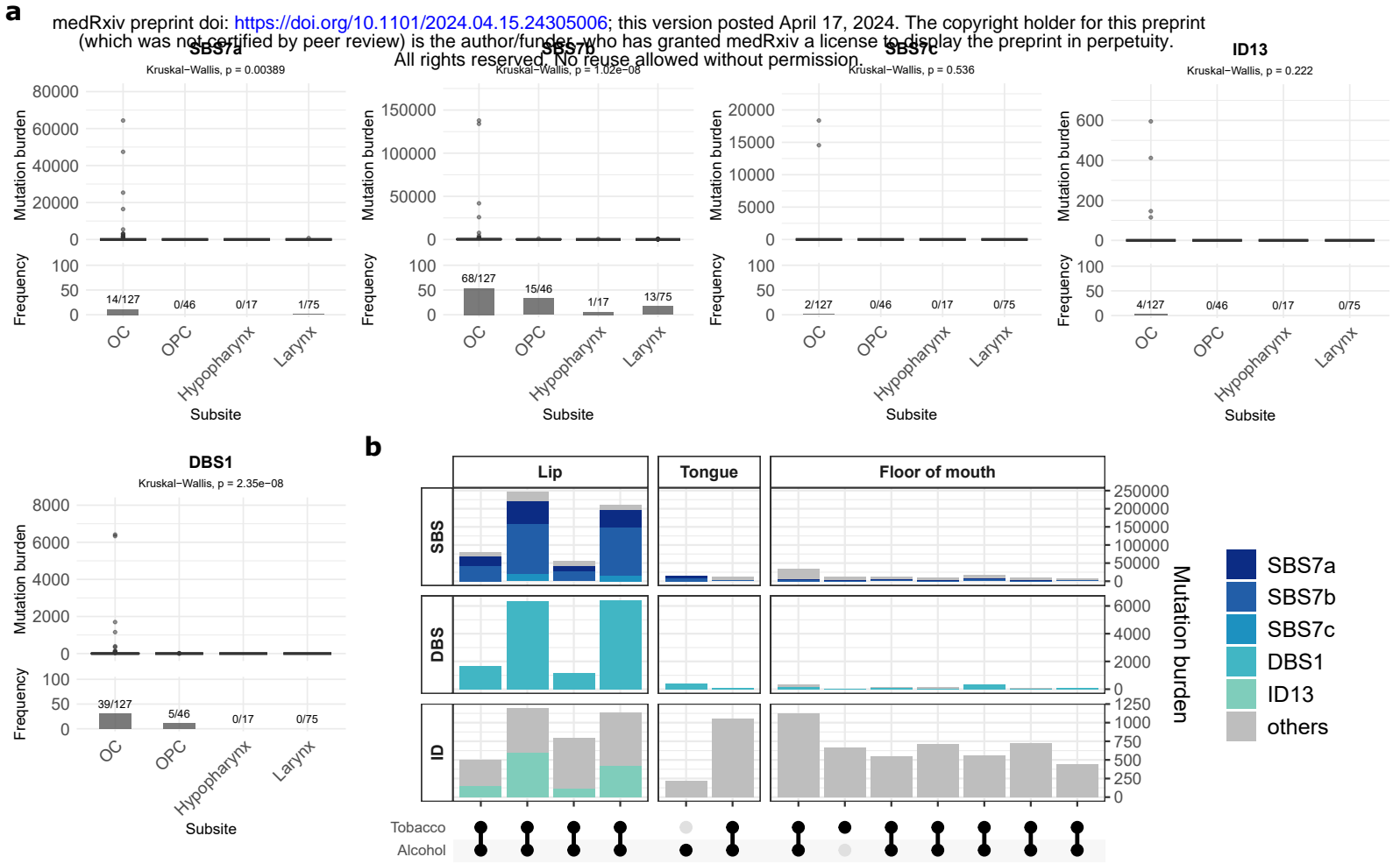
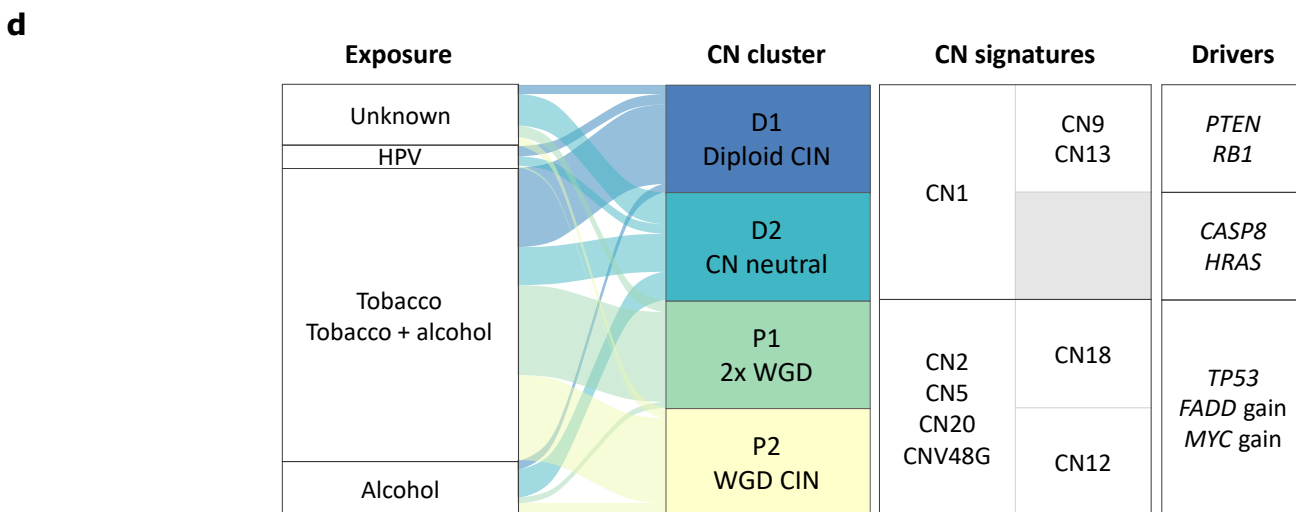
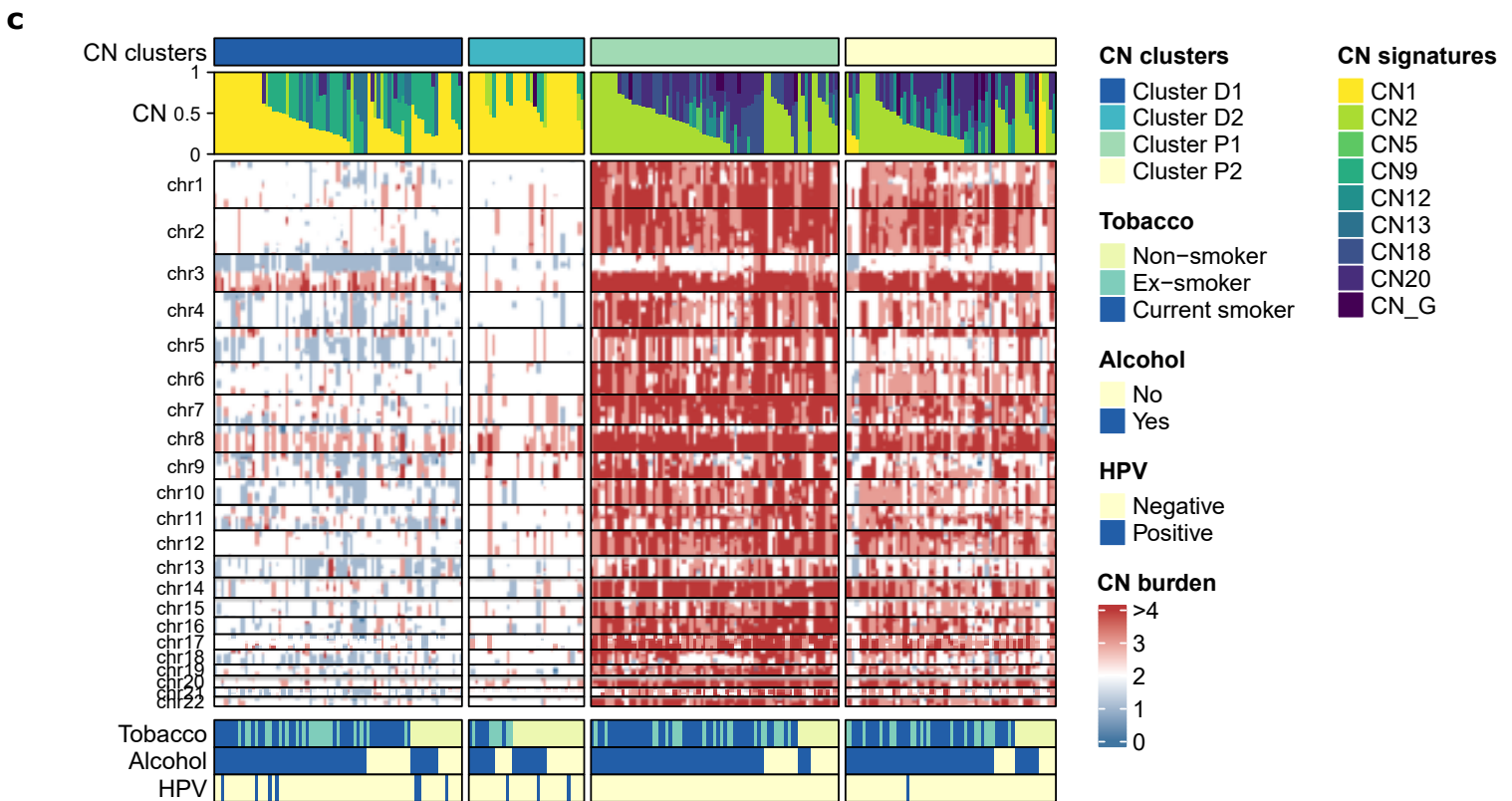
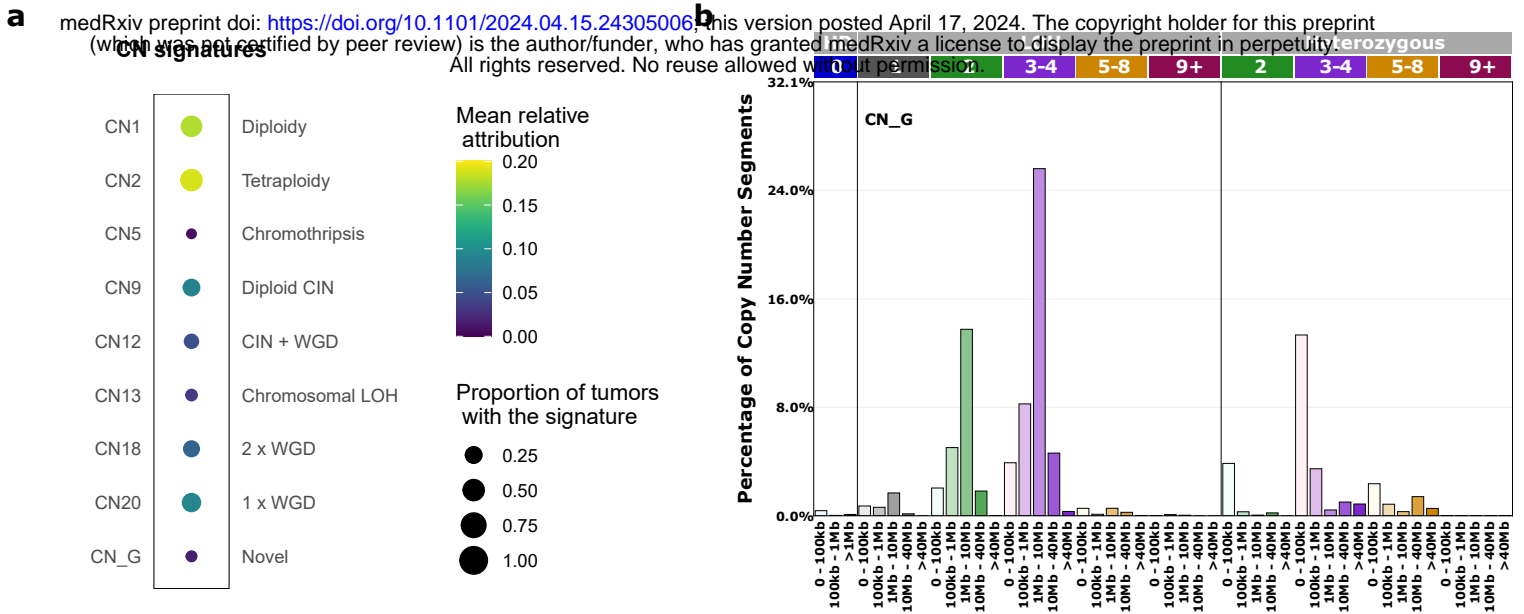
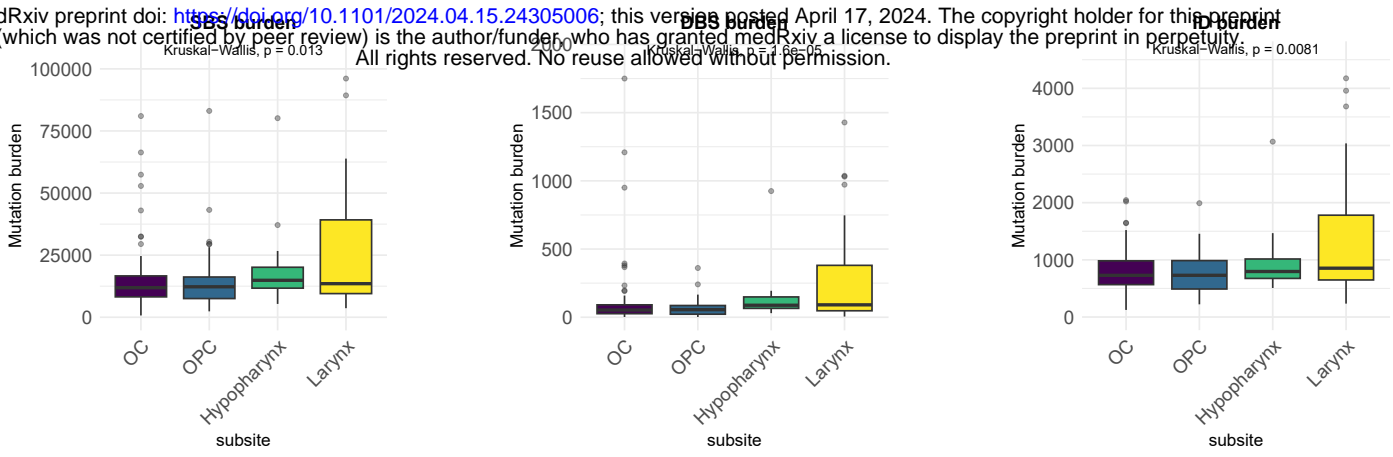


Figure 7

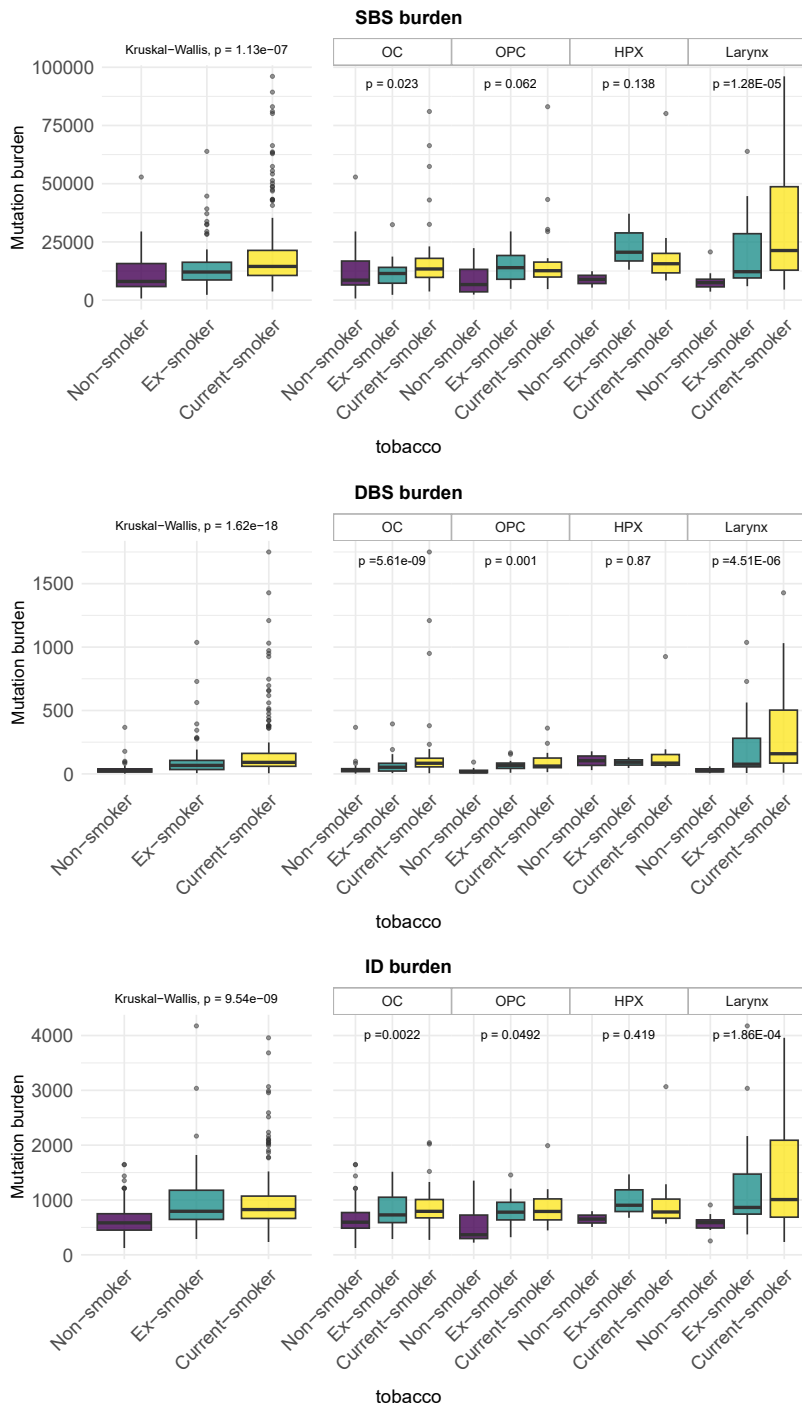


Extended Data Figure 1

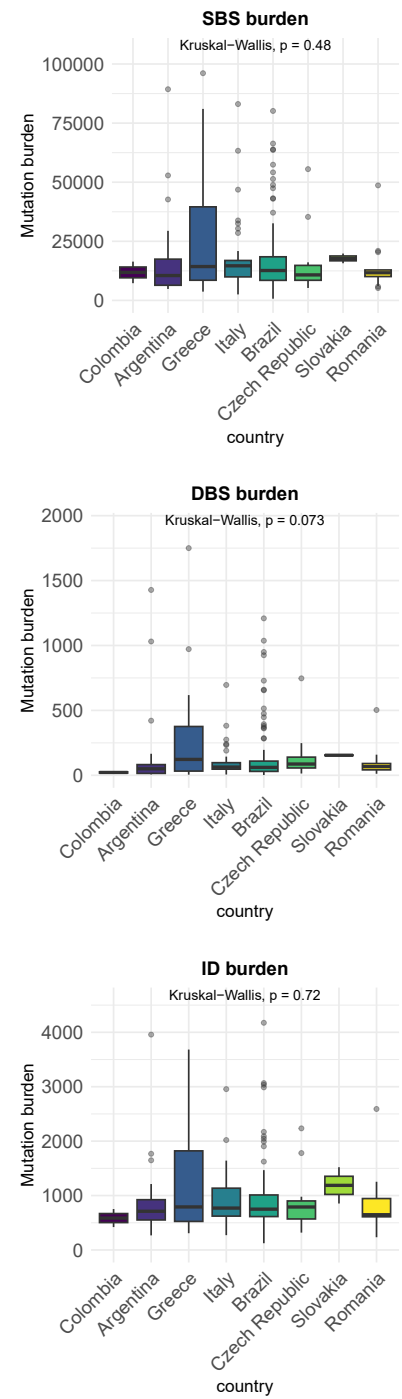
a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



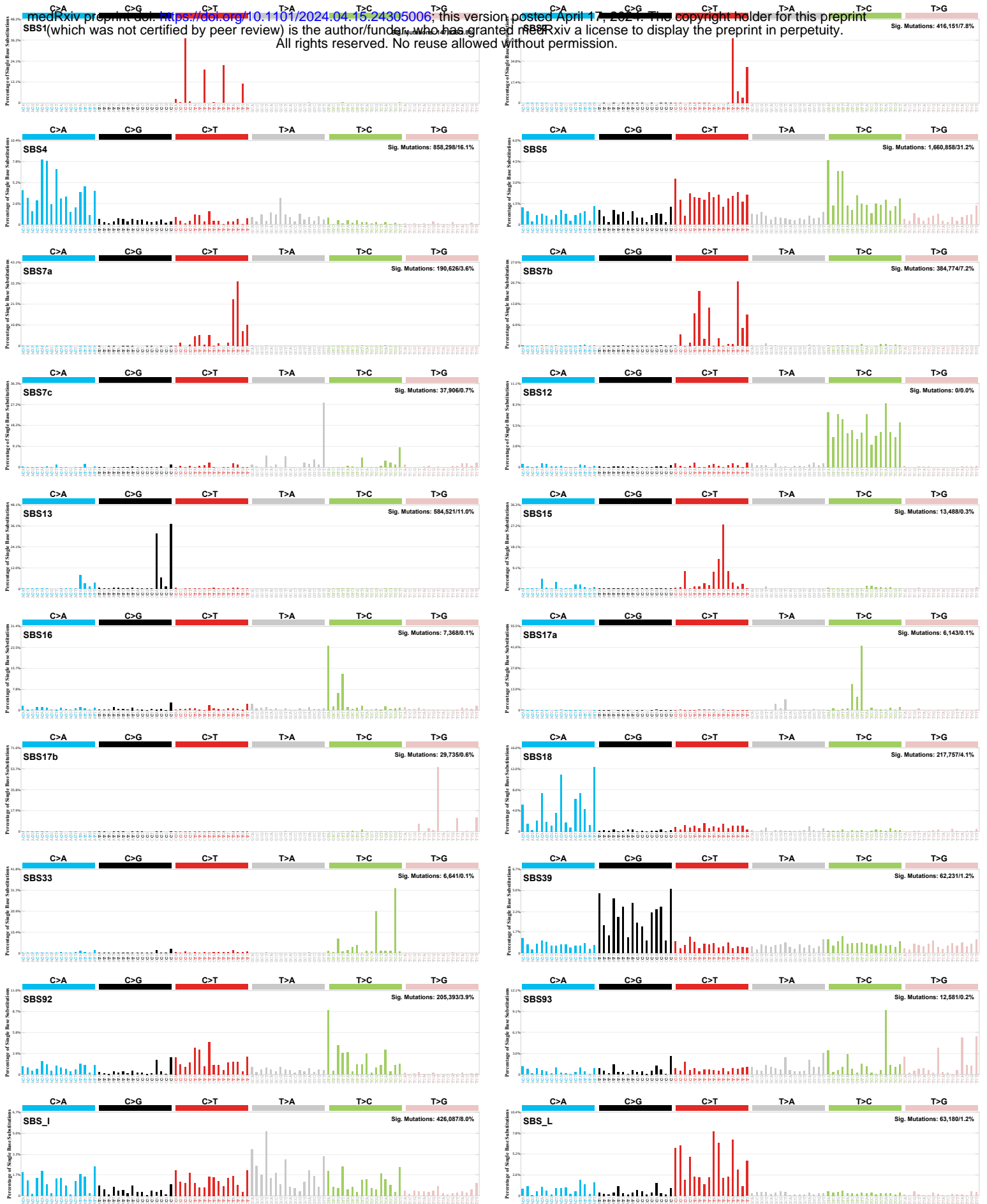
b



c

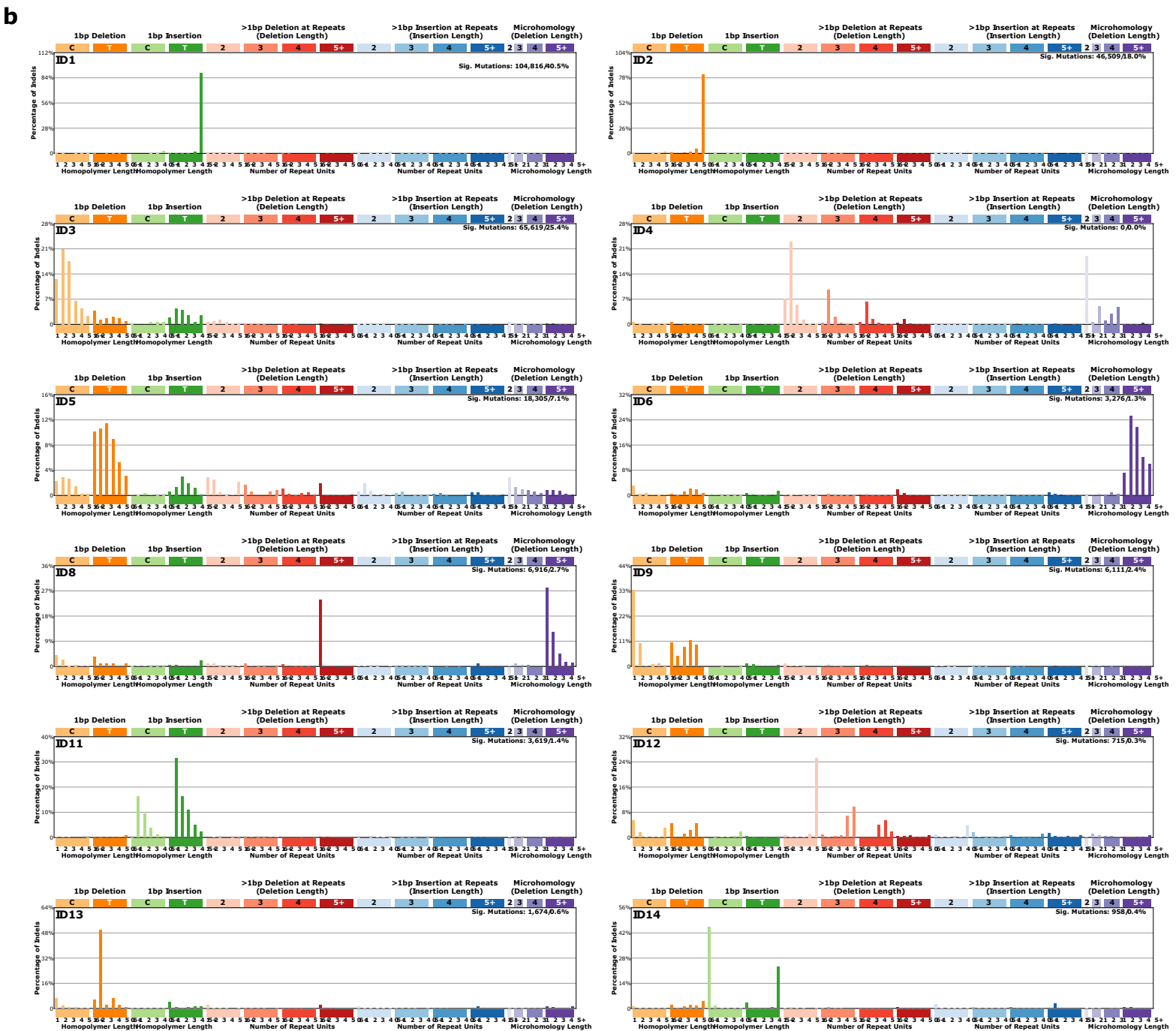
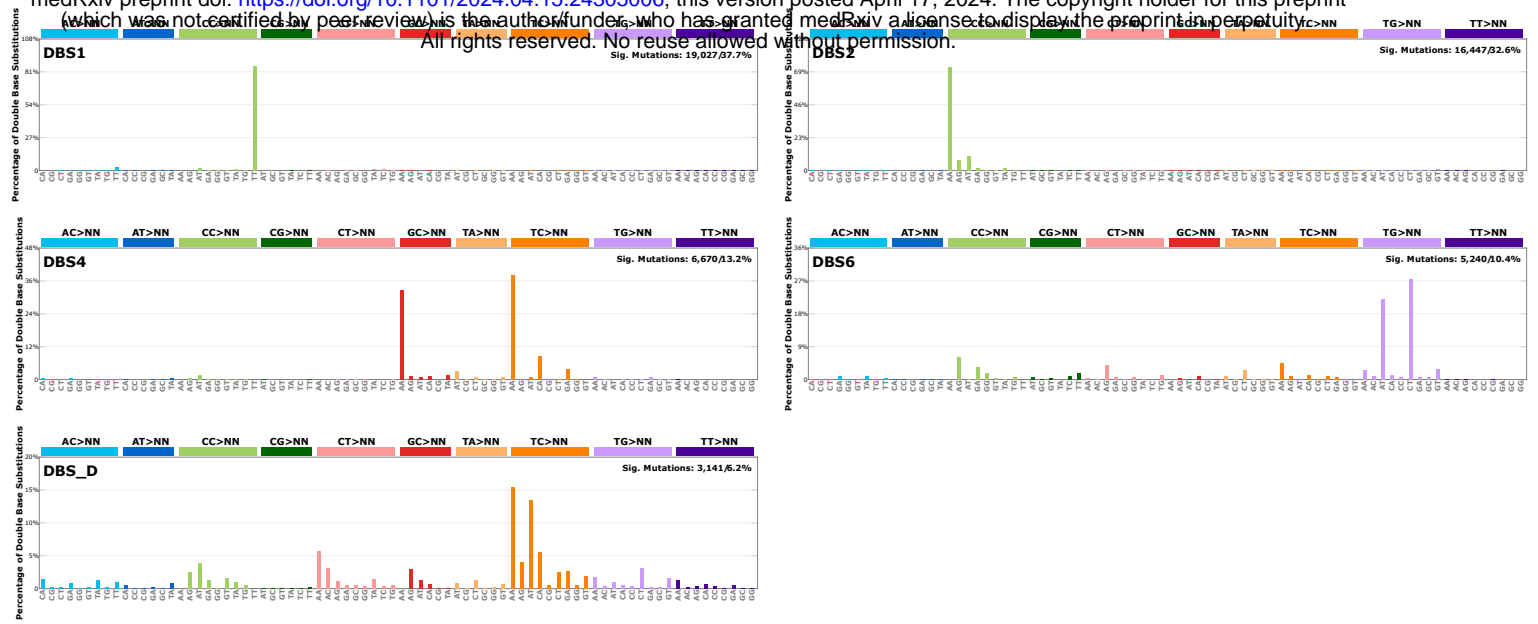


Extended Data Figure 2

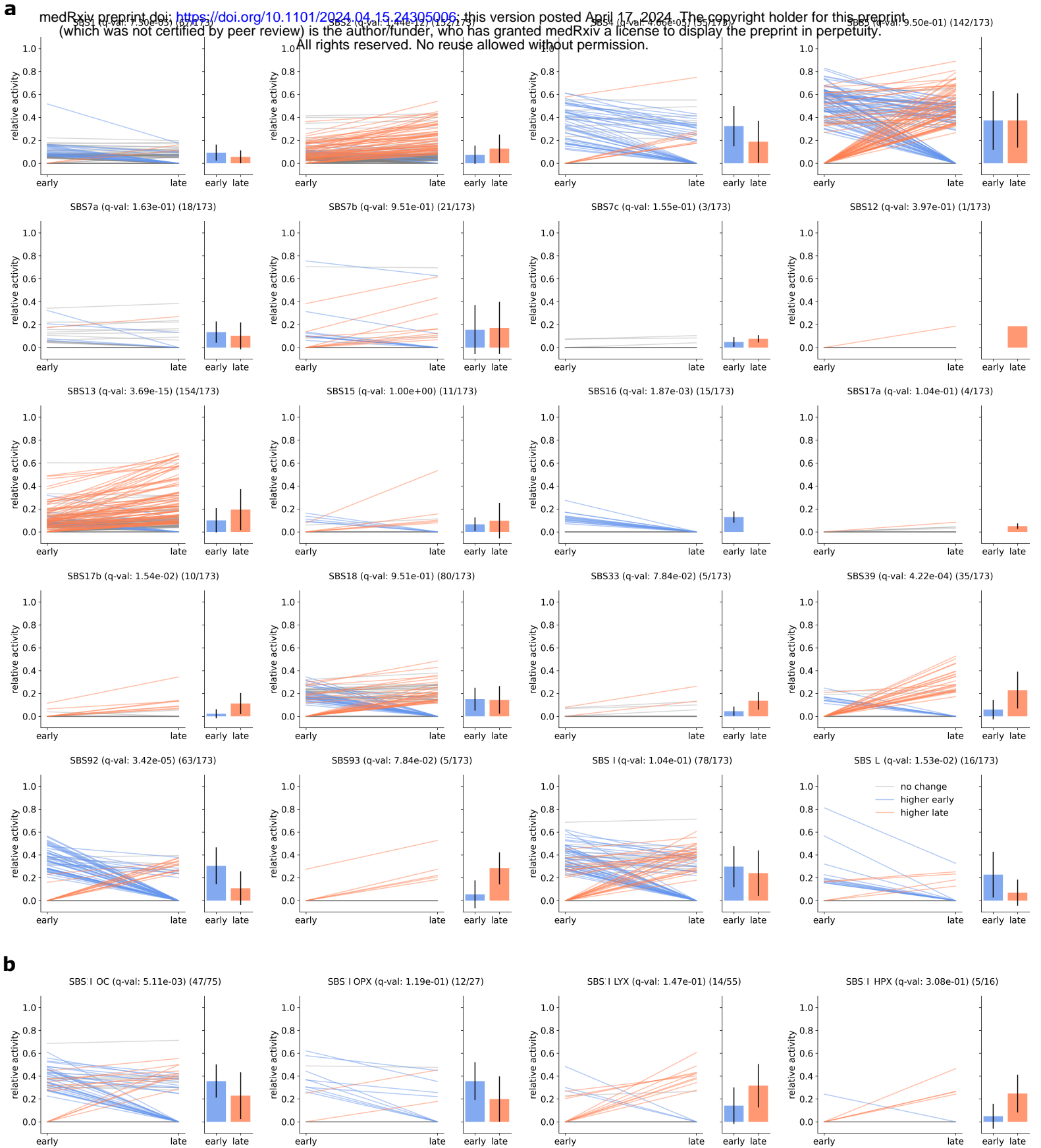


Extended Data Figure 3

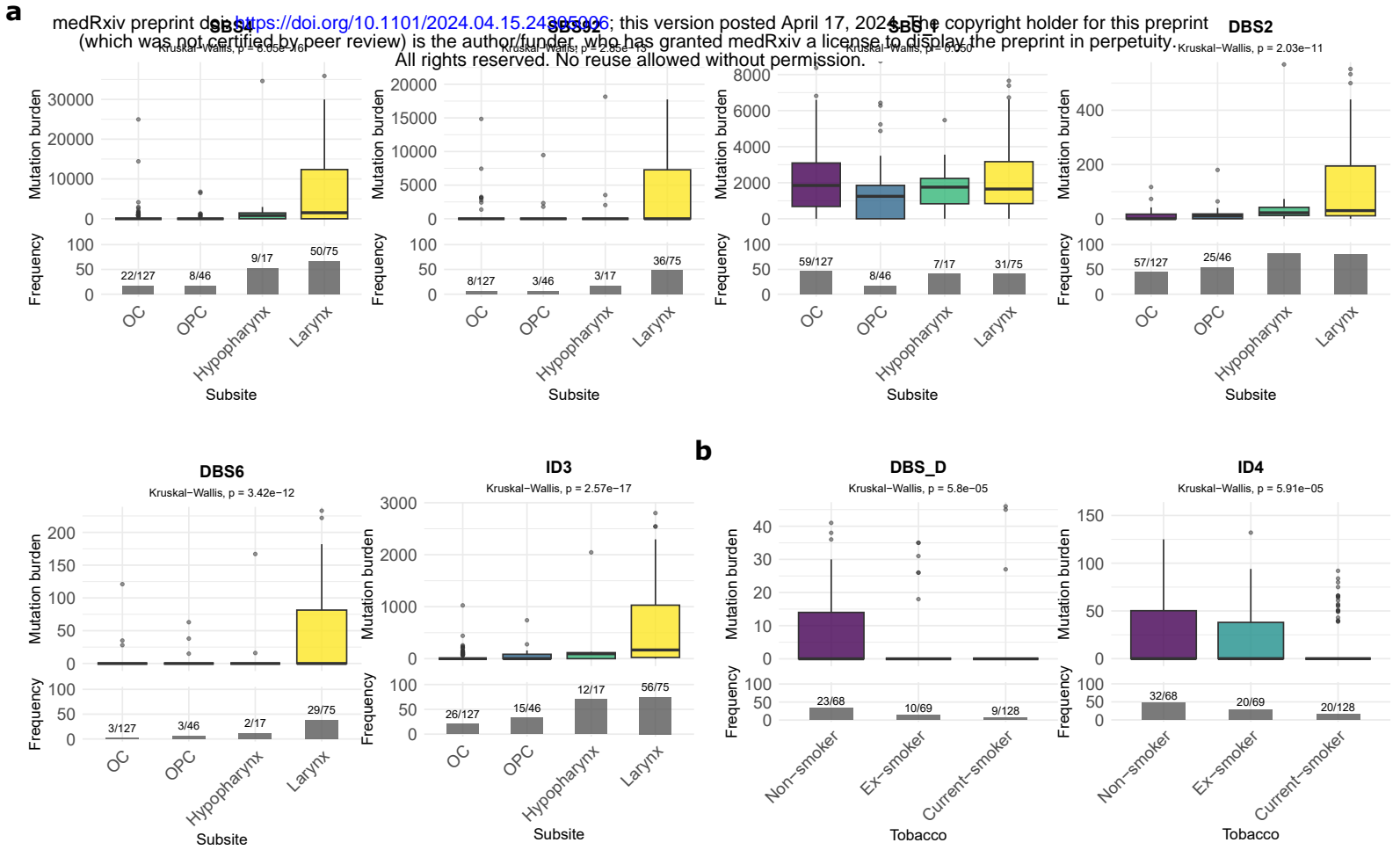
a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



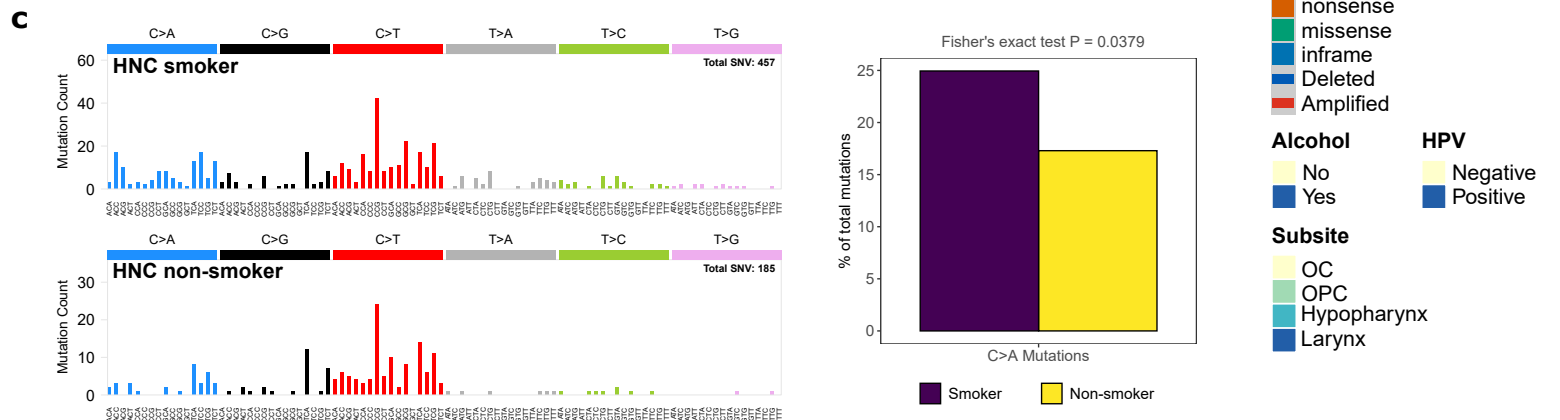
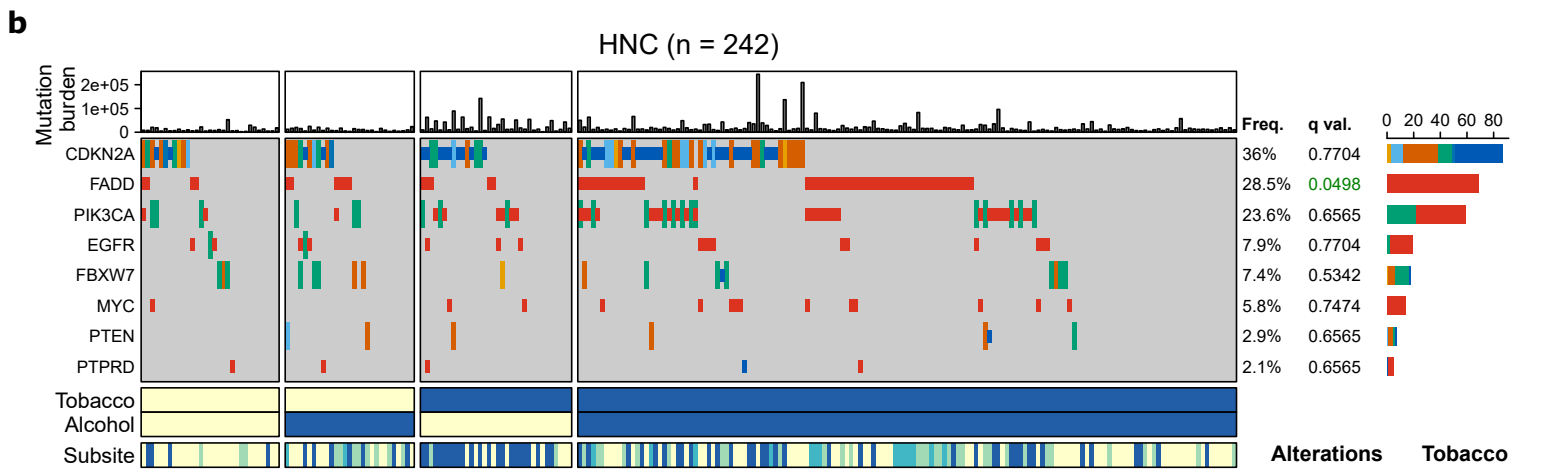
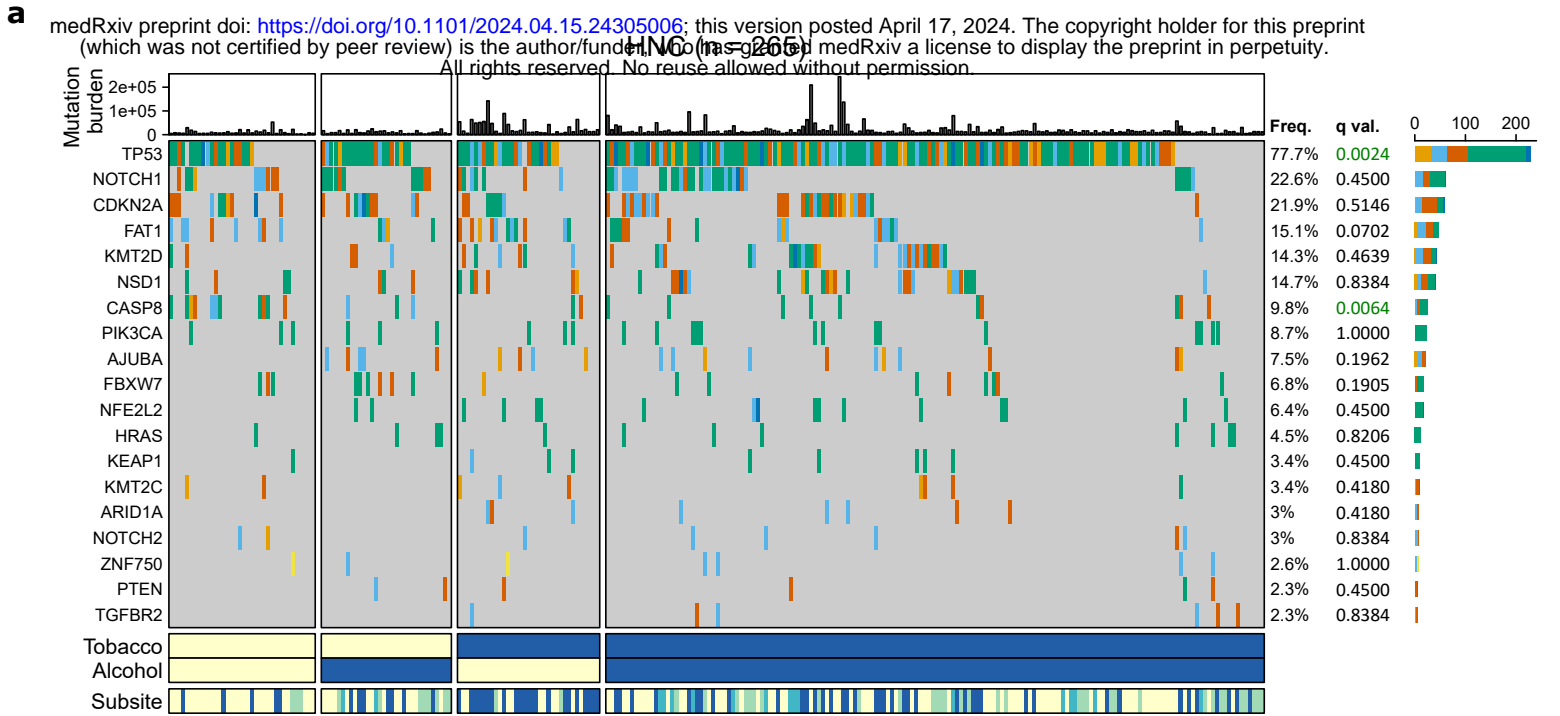
Extended Data Figure 4



Extended Data Figure 5



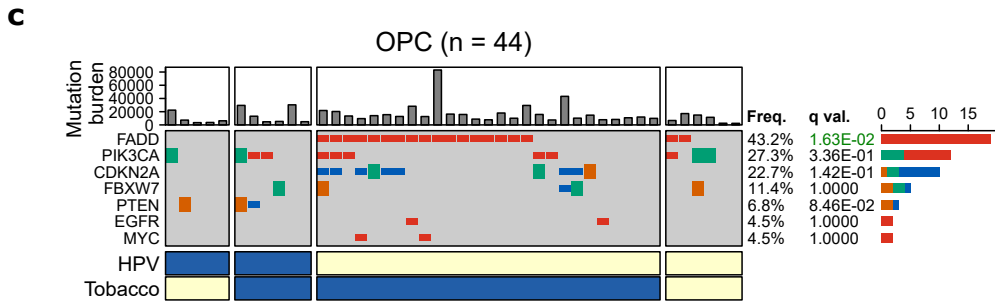
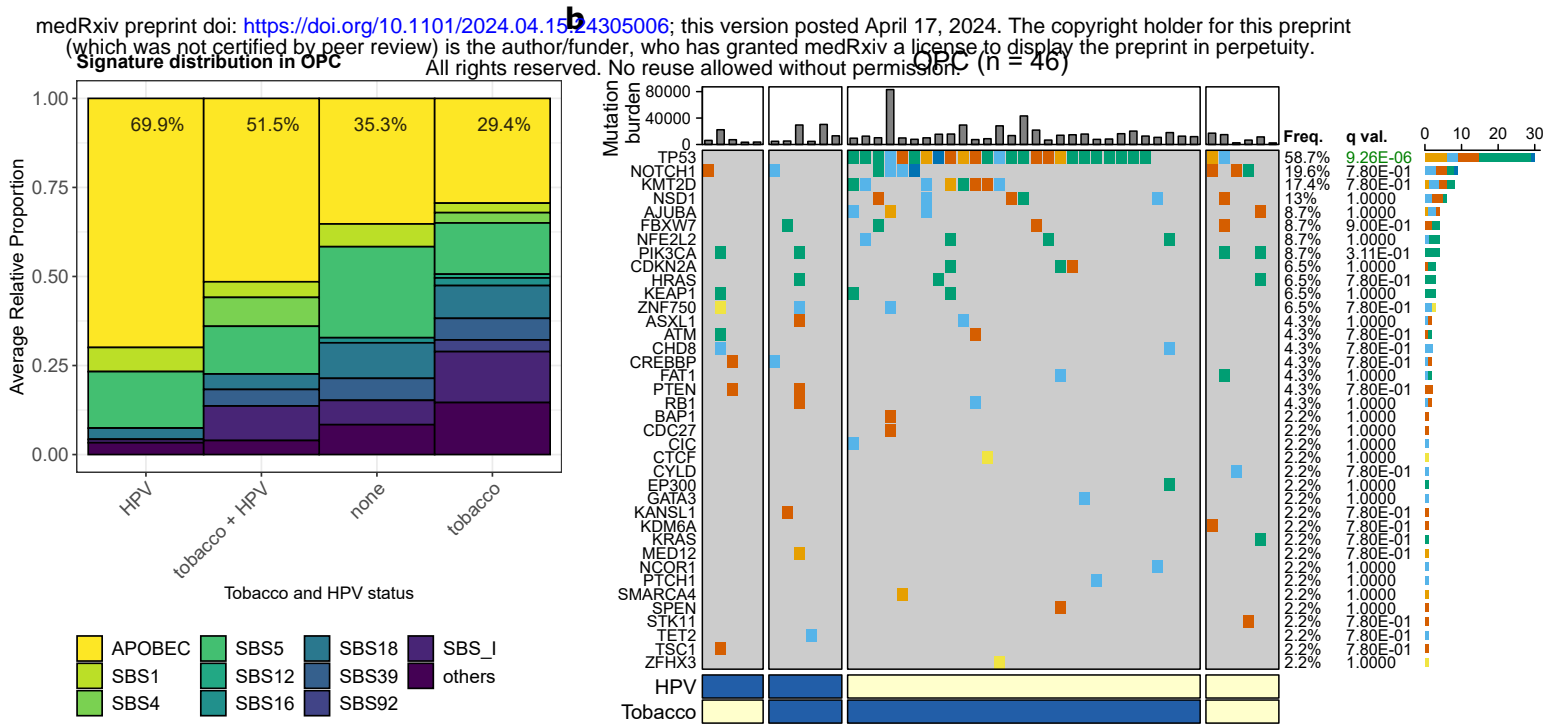
Extended Data Figure 6



- Alterations**
- ESS splice
 - frameshift
 - nonsense
 - missense
 - inframe
 - Deleted
 - Amplified
- Tobacco**
- No
 - Yes
- Alcohol**
- No
 - Yes
- HPV**
- Negative
 - Positive
- Subsite**
- OC
 - OPC
 - Hypopharynx
 - Larynx

Extended Data Figure 7

a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Alterations

- splicing
- frameshift
- nonsense
- missense
- inframe
- start lost
- Deleted
- Amplified

HPV

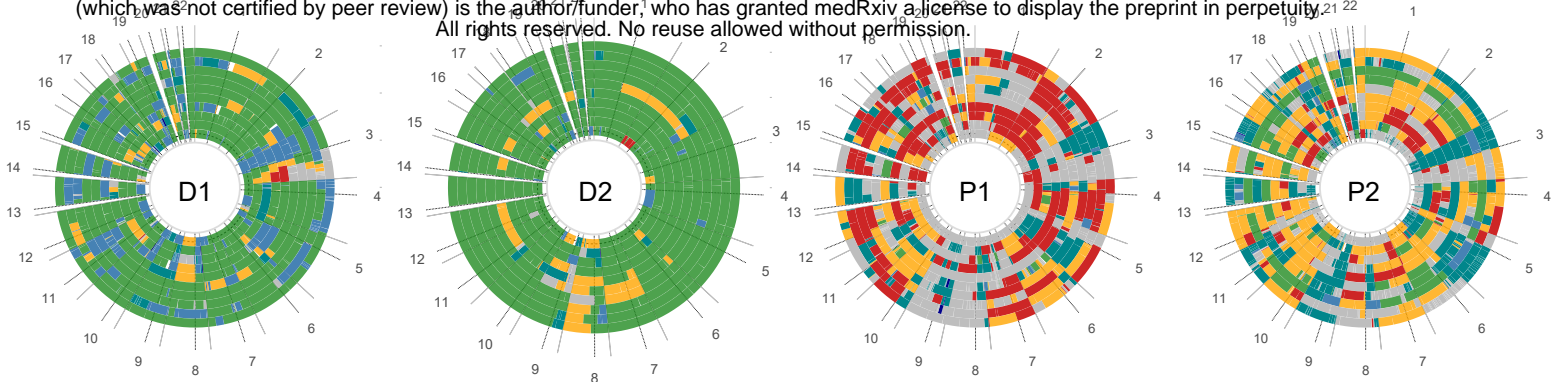
- Negative
- Positive

Tobacco

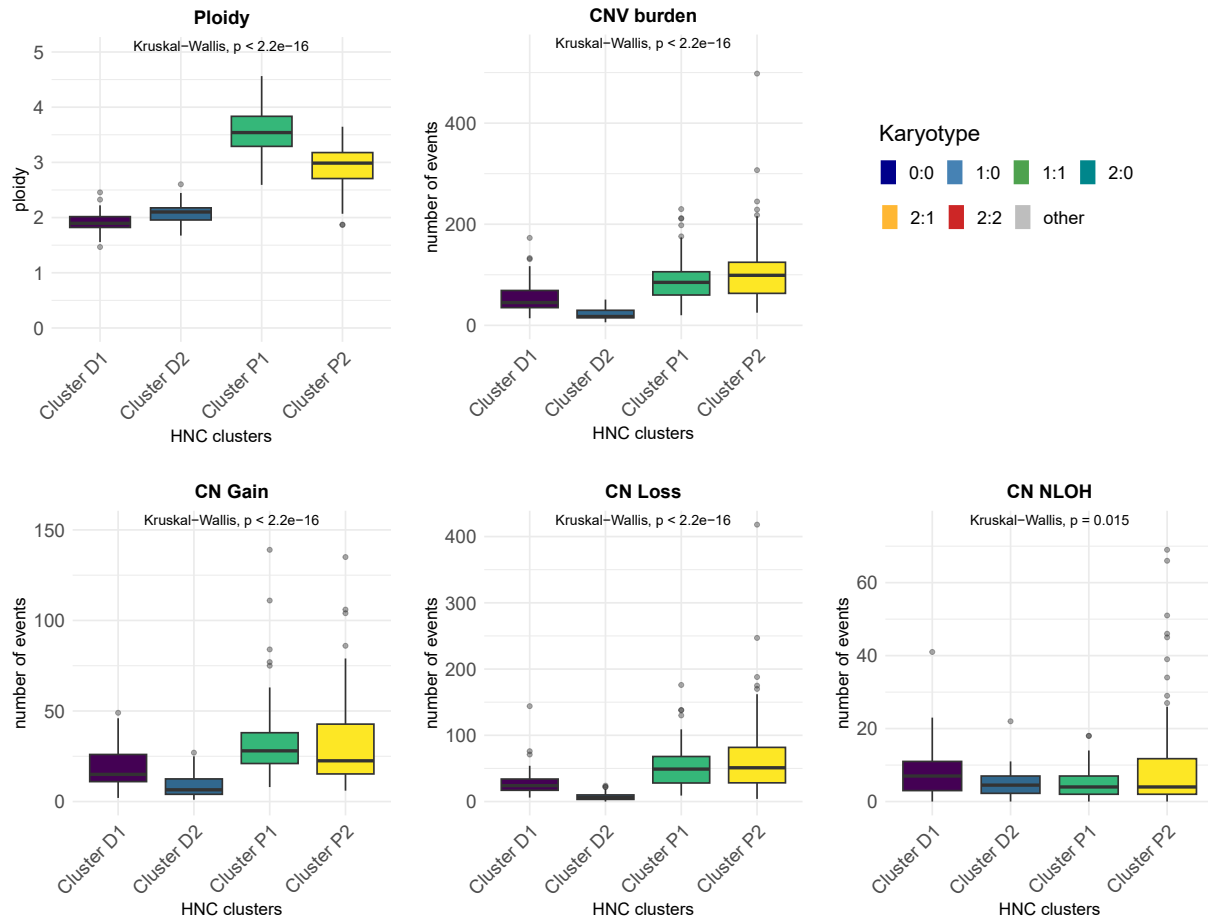
- No
- Yes

Extended Data Figure 8

a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

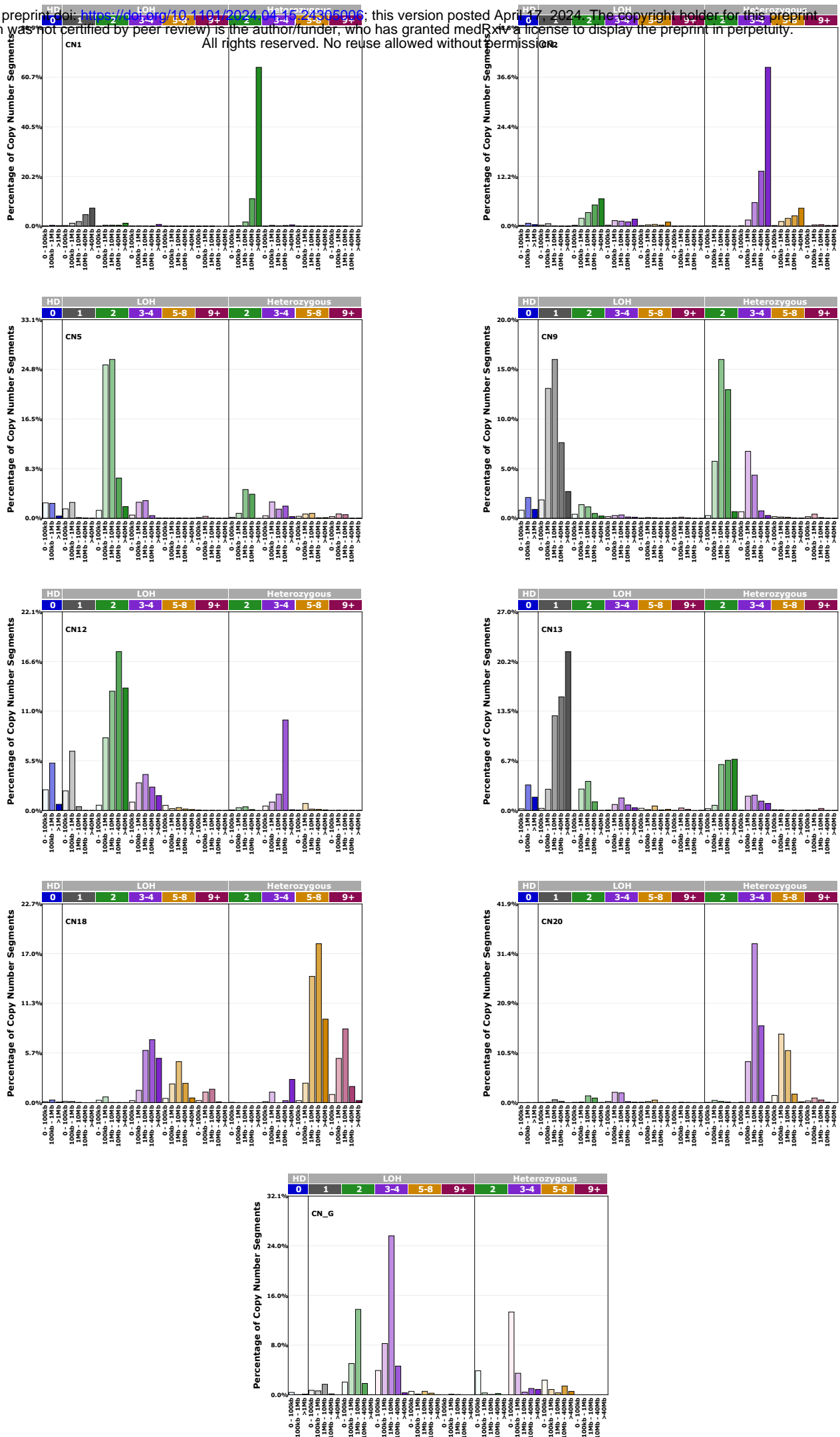


b



Extended Data Figure 9

medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305996>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Extended Data Figure 10

a medRxiv preprint doi: <https://doi.org/10.1101/2024.04.15.24305006>; this version posted April 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

