

Title: Somatic mosaicism in the buccal mucosa reflects lifestyle and germline risk factors for esophageal squamous cell carcinoma

Authors: Akira Yokoyama^{1*}, Koichi Watanabe², Yoshikage Inoue², Tomonori Hirano², Masashi Tamaoki¹, Kenshiro Hirohashi¹, Shun Kawaguchi², Yoshihiro Ishida^{2†}, Yasuhide Takeuchi³, Yo Kishimoto⁴, Soo Ki Kim⁵, Chikatoshi Katada¹, Yasuhito Nannya⁶, Hiroshi Seno⁷, Seishi Ogawa^{2,8,9}, Manabu Muto¹, Nobuyuki Kakiuchi^{2,7,10*}

Affiliations:

¹Department of Medical Oncology, Kyoto University Graduate School of Medicine, Kyoto, 606-8507, Japan.

²Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, 606-8501, Japan.

³Department of Diagnostic Pathology, Graduate School of Medicine, Kyoto University, Kyoto, 606-8507, Japan.

⁴Department of Otolaryngology-Head and Neck Surgery, Graduate School of Medicine, Kyoto University, Kyoto, 606-8507, Japan.

⁵Department of Gastroenterology, Kobe Asahi Hospital, Kobe, 653-0801, Japan.

⁶Division of Hematopoietic Disease Control, The Institute of Medical Science, The University of Tokyo, Tokyo, 108-8639, Japan.

⁷Department of Gastroenterology and Hepatology, Graduate School of Medicine, Kyoto University, Kyoto, 606-8507, Japan.

⁸Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, 606-8501, Japan.

⁹Department of Medicine, Centre for Hematology and Regenerative Medicine, Karolinska Institute, Stockholm, 171 77, Sweden.

¹⁰The Hakubi Center for Advanced Research, Kyoto University, Kyoto, 606-8501, Japan.

†Present address: Cancer, Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Hinxton, CB10 1RQ, UK.

*Corresponding authors. Email: yokoaki@kuhp.kyoto-u.ac.jp (Akira Yokoyama), kakiuchi@kuhp.kyoto-u.ac.jp (Nobuyuki Kakiuchi)

OVERLINE: CANCER

One Sentence Summary: Non-invasive sampling of the buccal mucosa can detect somatic mosaicism that marks age as well as lifestyle and germline risk factors for ESCC.

Abstract: Clones harboring cancer driver mutations can expand in normal tissues, known as somatic mosaicism, and can be influenced by age and environmental and germline factors. Somatic mosaicism in the blood predicts the risk of hematological malignancies; however, the relevance of somatic mosaicism to solid tumors remains unclear, in part due to limited sample availability. Lifestyle habits, including alcohol consumption and tobacco smoking, and pathogenic germline variants increase the risk of developing esophageal squamous cell carcinoma (ESCC). Because somatic mosaicism in the esophagus is known to be associated with aging and lifestyle habits, and considering the contiguity of squamous epithelium from the esophagus to the oral cavity, we non-invasively collected buccal mucosa samples from patients with and without ESCC using swabs of different sizes and conducted deep error-corrected sequencing of 26 cancer driver genes to obtain comprehensive landscapes of tissue remodeling by driver-mutant clones. We found that the number of mutations increased with drinking, but only in individuals with germline risks. Moreover, across positively selected genes in the buccal mucosa, mutations increased with age and smoking regardless of germline risks, whereas drinking affected only those with germline risks. The buccal mucosa of patients with ESCC was extensively remodeled, and models predicting the presence of ESCC demonstrated high accuracy with smaller swab sizes, possibly due to their higher sensitivity in detecting small mutant clones. In conclusion, we showed that buccal mucosal remodeling reflects lifestyle and germline risks, as well as age, which might be exploited for non-invasive risk assessment of ESCC.

INTRODUCTION

As we age and are exposed to environmental insults, individual cells in normal tissues accumulate somatic mutations, some of which confer a selective advantage, called driver mutations (1, 2). As tissue undergoes consecutive remodeling by proliferative cells, cells with a growth advantage are positively selected, and their descendant cells expand (1, 2). Our bodies consist of such expanded clones, a phenomenon known as somatic mosaicism (1, 3). Because driver mutations in normal tissues are largely shared with those found in the corresponding cancers (1, 4-17), driver-mutant clones in normal tissues are considered as the cell-of-origin for cancer. In the blood, the presence of clonal expansion, termed clonal hematopoiesis, is an established biomarker for the development of hematological malignancies (4, 5). Environmental factors, as well as germline variants, are known to affect the prevalence of clonal hematopoiesis (18-20). However, in solid organs, assessing cancer risks from tissue remodeling has not been established, although it has been reported that clones carrying cancer driver mutations pervasively remodel normal tissues in the elderly and patients with cancer in several organs, such as the esophagus (7, 21), bronchus (15), and bladder (8). Sampling of tissues in substantial quantities (beyond PBMCs) is not always feasible, and achieving a comprehensive landscape of tissue remodeling in solid organs from small biopsy areas remains an unsolved challenge.

Esophageal cancer is the sixth leading cause of cancer-related deaths worldwide, affecting approximately 604,100 individuals and causing 544,076 deaths annually (22). Of these cases, esophageal squamous cell carcinoma (ESCC) accounts for 85% (23). Despite the effectiveness of endoscopic screening tests for early detection, their cost and invasiveness pose challenges (24). Lifestyle habits, such as drinking and smoking, along with the presence of germline variants of *aldehyde dehydrogenase 2 (ALDH2)* and *alcohol dehydrogenase 1B*

(*ADH1B*), key genes in alcohol metabolism, are established risk factors of ESCC, with the combination of lifestyle and germline factors synergistically increasing ESCC risk (25-27). However, self-reported information on lifestyle habits and traits is often inaccurate, compromising risk stratification accuracy. Therefore, an objective, accurate patient stratification method is needed to facilitate screening. We previously demonstrated that driver-mutant clones remodel the esophageal epithelium with age and that this remodeling is accelerated in heavy drinkers and smokers, contributing to field cancerization (7, 28). This process is characterized by the expansion of cells influenced by carcinogenic alterations within an organ and is associated with an increased risk of ESCC (29). The squamous epithelium contiguously covers the esophagus, throat, and oral cavity. Heavy drinking and smoking increase cancer risk in these areas (30, 31), suggesting that driver-mutant clones may also remodel the buccal mucosa.

In this study, we non-invasively collected buccal mucosa samples using swabs of different sizes, followed by deep error-corrected targeted sequencing of 26 driver genes, to understand how known risk factors affect tissue remodeling in the buccal mucosa and to establish a screening tool for ESCC based on this remodeling.

RESULTS

Somatic mutations in ESCC driver genes within the buccal mucosa

A total of 222 patients (110 in the primary cohort and 112 in the additional cohort) (Fig. 1A), including 121 with ESCC and 101 without ESCC, were enrolled (table S1). To analyze the effect of lifestyle habits on buccal mucosal remodeling, we tried to match patients in both groups for age and heavy drinking habits, resulting in no significant differences in age, sex, heavy drinking, and heavy smoking habits ($p>0.05$) (table S2). Compared to subjects without ESCC, the ESCC

group had a higher prevalence of germline risk variants of *ALDH2* (84% vs. 32%, $p<0.01$) and *ADH1B* (22% vs. 8%, $p<0.01$). We used three swab sizes (20, 28, and 254 mm²) to obtain buccal mucosa samples because mutant clone sizes were unknown (fig. S1A). For each participant, samples were obtained from different parts of the central region of the buccal mucosa (solitary samples), whereas an additional three samples were obtained from a subset of participants ($n=26$) using the three swab sizes from the other sides (paired samples) (fig. S1, B and C). DNA was extracted from the swabs (fig. S1D), followed by library construction using unique molecular indices (UMIs) (see Materials and Methods). Targeted capture sequencing was performed on 26 driver genes known to be positively selected in ESCC or physiologically normal esophagus(7), as well as *ALDH2* and *ADH1B*, at a raw target depth of 20,000 $\times$ (fig. S1E). Consensus reads were then generated using both the top and bottom strands, each tagged with a unique set of UMIs. Samples with a consensus read depth of $\geq 1,800\times$ were included for further analysis ($n=707$ from 222 subjects) (fig. S1B), whereas samples with a depth $\geq 3,000\times$ were downsampled to 2,500 $\times$ to prevent heightened sensitivity in detecting mutations with low variant allele frequencies (VAFs). The consensus depths of each gene were from 956 $\times$ to 2219 $\times$ on average (table S3). The filter settings for mutation calling were determined using synthesized data generated by combining ESCC and germline data, both generated in the same manner as buccal mucosa samples, at ratios ranging from 0.2-1.0% (see Materials and Methods). This resulted in a sensitivity of 0.96 and a true positive rate of 0.96 for mutations with VAFs ≥ 0.0016 (fig. S1F). Among mutations detected in the buccal mucosa samples, digital droplet PCR experiments validated 104 out of 111 mutations (93.7%) in *TP53* with VAFs ranging from 0.0017-0.0108 (table S4) (see Materials and Methods).

Somatic mutations were detected in most samples [694 out of 707; median, 24 (0-163)]. In total, 23,537 mutations, including 22,136 nonsynonymous and 1,401 synonymous mutations, with VAFs ranging from 0.0016-0.12 (median, 0.0027), were detected (fig. S1G). Because mutational signatures in the buccal mucosal epithelium were previously unknown, we first employed NanoSeq (32) to detect mutations in the buccal swab samples in an unbiased manner ($n=16$) (table S5), followed by signature extraction. This resulted in the identification of COSMIC single base substitution (SBS) signatures (33): SBS1, SBS5, SBS16, and SBS40a (fig. S2A) (see Materials and Methods). SBS1, SBS5, and SBS40a have been reported to be associated with aging (34), whereas SBS16 has been associated with alcohol consumption (7, 35, 36), though this association is not yet considered conclusive. Mutations detected in each sample were then decomposed into these signatures (Fig. 1B and fig. S2B). The proportion of SBS5/SBS40a mutations—both relatively flat signatures that were combined—was similar between duplex sequencing and NanoSeq (fig. S2C). However, the proportions of SBS1 and SBS16 mutations were overrepresented and underrepresented, respectively, in the duplex sequencing data (fig. S2C). This discrepancy may partly be explained by the higher proportion of NCG trinucleotides—C>T mutations in these contexts are characteristic of SBS1—and the lower proportion of ATA/ATT/ATG trinucleotides—T>C mutations in these contexts are characteristic of SBS16—in the 26 genes compared to the genome (fig. S2D). Although 94% of detected mutations were attributable to aging signatures, the participants' age had only a modest effect on the number of mutations (fig. S2E). Instead, the average VAF of detected mutations showed a significant correlation after excluding outliers with high average VAFs ($p=9.7\times 10^{-38}$), suggesting that the detection of mutant clones was influenced by their clone size (fig. S2F). Next, we fitted a linear mixed-effects (LME) model to assess the impact of known risk factors—age, alcohol

consumption, smoking intensity, and the presence of germline risk variants—on the number of mutations (Fig. 1C). Germline risk was the most significant factor ($p < 2.0 \times 10^{-16}$), followed by alcohol consumption, prompting speculation about a potential interaction between the two. We then examined the effect of alcohol consumption on the number of mutations among participants with and without germline risk variants. Alcohol consumption was positively correlated with the number of mutations in subjects carrying germline risk variants ($p < 0.05$), whereas the correlation in germline low-risk participants was not significant ($p = 0.29$) (Fig. 1D). Although signature decomposition might produce substantial variance, a similar trend was observed when the analysis was limited to the number of SBS16 mutations (Fig. 1E). The correlation between alcohol consumption and the number of SBS16 mutations in germline high-risk participants remained significant after excluding outliers with high SBS16 mutations ($p = 6.8 \times 10^{-3}$), suggesting that the effect of alcohol consumption varies depending on germline risk status.

Clonal expansion driven by ESCC driver mutations in buccal mucosa

Among the clones detected using 254 mm² swabs, the largest clone was 48.38 mm² and carried a nonsense mutation in *protein phosphatase, Mg²⁺/Mn²⁺ dependent 1D (PPM1D)*, whereas 90% were ≤ 3.0 mm² in size (figs. S1G and S3). In contrast, the largest clones detected using 20 mm² and 28 mm² swabs were 4.74 mm² and 5.02 mm² in size, respectively, and 90% were ≤ 0.30 mm² and ≤ 0.41 mm² (median, 0.10 mm² and 0.15 mm²), respectively (figs. S1G and S3), which were undetectable using a 254 mm² swab due to the lower limit of detectable VAF (≥ 0.0016 or 0.81 mm²). To validate the clone sizes detected using swabs, we enrolled a 58-year-old patient who underwent resection of a buccal mucosal carcinoma. Normal mucosa adjacent to the resected cancer was analyzed (fig. S4, A to C). After separating the epithelium from the stroma, 0.2 mm² samples ($n=40$) were collected in 1-mm lattices using punch biopsy, followed by whole-exome

sequencing (see Materials and Methods). A clone carrying the *TP53* mutation was observed in four samples (about 4 mm²) (fig. S4A), whereas most mutations were confined to individual samples. The median size of the mutant clones was 0.1 mm² (VAF of 0.25), which was above the lower limit of detectable clone size using 28 mm² and 20 mm² swabs (0.064 mm² and 0.09 mm², respectively). Together, these results indicated that the buccal mucosa comprises numerous clones around 0.1-0.3 mm² in size, with some clones occasionally expanding to sizes as large as tens of mm². Therefore, the areas of the three swab sizes were adequate to collect enough mutant clones to provide a comprehensive landscape of buccal mucosal remodeling.

The number and sum of VAFs of detected mutations were compared between samples obtained from the same individuals using different swab sizes, revealing substantial concordance and a trend toward higher values in samples collected with smaller swabs (fig. S5A). This trend may partly be explained by the higher detection sensitivity of smaller clones when using smaller swabs. To assess the uniformity of buccal mucosal remodeling, we applied linear regression to compare numbers and sums of VAFs of detected mutations between paired samples obtained from the left and right buccal mucosa of the same individuals, demonstrating a high concordance (fig. S5B). We also compared tissue remodeling between the buccal and esophageal mucosa by enrolling an additional 10 cases, five with ESCC and five without (table S6). From each individual, both esophageal and buccal samples were collected, followed by duplex sequencing (see Materials and Methods). Although the number of mutations showed weak correlations, the sum of VAFs exhibited modest correlations between the buccal and esophageal mucosa (fig. S5C). The lower concordance may partly be explained by the size of the esophageal area analyzed. Duplex sequencing of a 10 mm² area of the esophagus detected a substantial number of small mutant clones, even in individuals without ESCC. In addition, the presence of relatively

large clones in the esophagus hindered capturing the full landscape of tissue remodeling at this area size. However, obtaining larger samples from the esophagus was impractical due to the invasiveness of the procedure, highlighting the advantage of buccal swab sampling.

Notch receptor 1 (NOTCH1) was the most frequently mutated gene, followed by *TP53* and *checkpoint kinase 2 (CHEK2)*, with proportions of 41.6%, 28.1%, and 5.4%, respectively (Fig. 2A). To investigate the mode of natural selection in the buccal mucosa, a ratio of non-synonymous to synonymous substitution (dN/dS) analysis was conducted using the R package *dndscv* (37) (table S7). Among the 26 genes analyzed, six genes, *TP53*, *CHEK2*, *NOTCH1*, *NOTCH2*, *FAT atypical cadherin 1 (FAT1)*, and *ajuba LIM protein (AJUBA)*, exhibited dN/dS >1 ($q < 0.1$) for both missense and protein-truncating mutations, indicating positive selection of mutations in these genes (Fig. 2B). The mutation distribution of each gene largely overlapped between the buccal mucosa, physiologically normal esophagus, and ESCC samples (Fig. 2C and fig. S6A), whereas six genes, *PPM1D*, *cullin 3 (CUL3)*, *AT-rich interaction domain 1A (ARID1A)*, *zinc finger protein 750 (ZNF750)*, *ZFP36 ring finger protein like 2 (ZFP36L2)*, and *cyclin dependent kinase inhibitor 2A (CDKN2A)*, showed positive selection for either missense or truncating mutations. In contrast, three genes, *NOTCH3*, *patched 1 (PTCH1)*, and *plexin B2 (PLXNB2)* showed evidence of negative selection for missense mutations (dN/dS <1, $q < 0.1$) (fig. S6B). Hereafter, our analysis focused on driver mutations, defined as positively selected non-synonymous mutations in *TP53*, *CHEK2*, *NOTCH1*, *NOTCH2*, *FAT1*, or *AJUBA*; or truncating mutations in *PPM1D*, *ARID1A*, *ZNF750*, *ZFP36L2*, or *CDKN2A*; or missense mutations in *CUL3*.

Relationship between mutations in buccal mucosa and ESCC risk factors

We analyzed the correlations between known risk factors for ESCC, including age, alcohol consumption, smoking intensity, and germline variants of *ALDH2* and *ADH1B*, and driver mutations in each positively selected gene in terms of the number of driver mutations and the sum of their VAFs using an LME model with random effects of swab size (Fig. 3A). Among the risk factors, germline risk variants of *ALDH2* exhibited the highest number of positive correlations, showing correlations with seven genes. In addition to germline risk variants of *ALDH2*, driver mutations in *TP53*, *CHEK2*, *NOTCH1*, *FAT1*, and *PPM1D* also positively correlated with alcohol consumption. The impact of alcohol consumption on the number and sum of VAFs of driver mutations in each of the five genes were analyzed and stratified by the presence and absence of the germline risk variant of *ALDH2* (Fig. 3B and fig. S7, A to D). Positive correlations were observed only among individuals with risk variants, indicating that alcohol consumption and germline risk variants of *ALDH2* combinatorially contribute to the expansion of clones carrying mutations in these genes. In contrast, among germline low-risk individuals, driver mutations were correlated more with smoking intensity than with alcohol consumption (Fig. 3C). Age was correlated with mutations in several driver genes, suggesting age-related clonal expansion in the buccal mucosa, consistent previous reports (7).

Correlation of buccal mucosal remodeling with lifestyle and ESCC characteristics

To quantify buccal mucosal remodeling by positively selected clones, the total number of driver mutations and the sum of their VAFs were calculated for each sample. Compared to individuals without germline risk variants, those with germline variants exhibited a greater extent of remodeling ($p<0.05$) (fig. S8A). The correlation of buccal mucosal remodeling to drinking, smoking, and age was analyzed using an LME model. Alcohol consumption was positively correlated only in individuals with germline variants ($p<0.05$) (Fig. 4, A and B). In contrast,

smoking intensity and age showed positive correlations regardless of germline variants (Fig. 4, A and B and fig. S8B). We next investigated the relationships between buccal mucosal remodeling and ESCC and found that individuals with extensive buccal mucosal remodeling also had higher rates of ESCC than those with less extensive remodeling (Fig. 4C). Clinical characteristics related to ESCC were also analyzed. Compared to those without ESCC, patients with early-stage or advanced ESCC displayed significantly higher numbers and sums of VAFs of driver mutations ($p<0.05$) (Fig. 4D). Head and neck squamous cell carcinoma (HNSCC) often occurs concurrently or metachronously in patients with ESCC. Patients with both ESCC and HNSCC showed significantly higher numbers and sums of VAFs of driver mutations than those with ESCC only ($p<0.05$) (Fig. 4E). An elevated mean corpuscular volume (MCV) has been found to correlate with the metachronous ESCC (38). Patients with $MCV \geq 106$ fL exhibited more extensive remodeling than those with $MCV < 106$ fL ($p<0.05$) (fig. S8C). Lugol voiding lesion (LVL) grade C is associated with a higher risk of metachronous ESCC and HNSCC than LVL grades A/B (31). The numbers and sums of VAFs of driver mutations were significantly higher in individuals with grade C than in those with grades A/B ($p<0.05$) (fig. S8D).

Buccal mucosal remodeling as a marker of ESCC presence

We next sought to develop a screening tool for ESCC based on buccal mucosal swab data (see Materials and Methods). We constructed models using data from the primary cohort and then applied these models to data from the additional cohort for external validation (Fig. 1A). We first identified a set of driver genes, mutations in which were significantly correlated with the presence of ESCC through univariate analyses ($p<0.05$), for each of the three swab types. Then, multivariable logistic regression was applied to data from a randomly selected 75% of samples in the primary cohort, with variable selection performed using the Least Absolute Shrinkage and

adapted Selection Operator (LASSO) method to eliminate confounders and prevent model overfitting. Variables selected in $\geq 80\%$ of ten thousand trials were used for constructing models (Fig. 5A). The number of mutations in *TP53*, *CUL3*, and *NOTCH1* were selected for the 20 mm² swab, whereas the mutation number of *TP53* and *FAT1* and those of *TP53* and *CHEK2* were selected for the 28 mm² and 254 mm² swabs, respectively. To evaluate the model performance, we randomly selected 75% of the samples to train the model based on the selected variables, and the concordance indices (c-indices) were calculated using the remaining 25% of samples; this procedure was repeated ten thousand times, resulting in high c-indices (median 0.82-0.91) (fig. S9A). Because the clinical characteristics of individuals in the primary and additional cohorts were not identical (table S2), we employed propensity score matching to select participant pairs with similar characteristics before model construction, aiming to reduce selection bias (fig. S9B). A logistic regression model, incorporating the selected variables, was constructed based on the matched data from the primary cohort and then applied to the matched data from the additional cohort. ROC curves for the 20 and 28 mm² swabs demonstrated high concordance with high AUC values between the two cohorts, indicating the robustness of the models, while the 254 mm² swab showed less concordance, partly due to its lower sensitivity in detecting mutant clones with this size of swab (AUC values: 20 mm², 0.84; 28 mm², 0.85; 254 mm², 0.77 in the primary cohort; and 20 mm², 0.83; 28 mm², 0.85; 254 mm², 0.85 in the additional cohort) (Fig. 5B). A threshold was determined as the point closest to (0, 1) on the ROC curve for the primary cohort, resulting in comparable sensitivity and specificity between the two cohorts (fig. S9C). We further examined the utility of these models among individuals with and without germline risk variants (Fig. 5C). AUC values were slightly lower for those with low-risk germline variants (AUC, 0.73 – 0.79) compared to those with high-risk germline variants (AUC, 0.83-0.85). This

may partly be explained by the inclusion of *TP53* mutations in the models, which showed the most significant correlation with the presence of ESCC in individuals both with and without germline risk variants ($p < 0.05$) (fig. S9D). Because the extent of buccal mucosal remodeling was more extensive in patients with both ESCC and HNSCC (Fig. 4E), these patients exhibited higher probabilities of ESCC presence, calculated by the models, in both the primary and additional cohorts (Fig. 5D and fig. S9E).

DISCUSSION

In this study, we used non-invasive swabs ranging from 20 to 254 mm² to obtain samples from the that demonstrated landscapes of buccal mucosal remodeling and showed that the remodeling by mutant clones reflected the extent of lifestyle risk factors, including drinking and smoking, as well as age and germline risk variants of *ALDH2* and *ADH1B*. In addition, we established that somatic mosaicism in the buccal mucosa can be used as an objective marker for the presence of ESCC.

Among the 26 genes positively selected in ESCC or the physiologically normal esophagus (7), 12 genes were positively selected in the buccal mucosa, whereas three genes were negatively selected. This suggests the presence of common and differential selection mechanisms between cancer and normal tissues, in line with previous reports (7, 10, 39). Tissue remodeling by driver-mutant clones, as well as the accumulation of SBS16 mutations, were accelerated by alcohol consumption more in those with germline risk variants, suggesting that germline and environmental factors jointly influence tissue remodeling. This may have relevance to the recently reported observation that germline risk variants of *BRCA1 DNA repair associated* (*BRCA1*) and *BRCA2* increase gastric cancer risk only in individuals with chronic *Helicobacter pylori* infection (40). In addition, exposure to alcohol did not accelerate clonal expansion among

those without germline risk variants, consistent with the low incidence of ESCC in countries where the prevalence of the germline risk variants is low (29). We constructed diagnostic models for ESCC, demonstrating that models incorporating mutations in several genes, including *TP53* and *CHEK2*, accurately diagnosed the presence of ESCC. These two genes belong to the DNA damage response pathway (41, 42), suggesting that functional deficiencies in this pathway may play a crucial role in ESCC tumorigenesis. Regarding the low specificity of our models, false positives—individuals who did not have ESCC at the time of enrollment—may be at high risk of developing ESCC, because their tissues showed extensive remodeling by mutant clones correlated with ESCC presence. This will need to be further evaluated in future studies.

Risk stratification for ESCC based on lifestyle habits and alcohol flushing reaction has been in practical use, with examples including the health risk appraisal (HRA) model (43) and the alcohol use disorders identification test (AUDIT) (44). However, accuracy is not high, in part due to the challenges of precisely quantifying drinking and smoking habits and estimating germline risk from traits like alcohol flushing reaction (43). It has been reported that *TP53* mutations were more frequently detected in the esophageal epithelium of patients with high LVL grades (31). However, due to the invasiveness of endoscopic examinations, mutation analysis of the esophageal mucosa has not been utilized for assessing ESCC risks. The optical properties of the buccal mucosa, measured using non-invasive reflectance spectroscopy, have been reported to reflect field cancerization (45), but the reported sensitivity (0.667) was lower than our method, although a direct comparison was not made in this study.

There are several limitations to this study that warrant consideration. Clinical information, particularly regarding alcohol consumption and smoking intensity, might be inaccurate, which could compromise the correlation analysis between mutant clones in the

buccal mucosa and known risk factors. Because this was a case-control study, some bias in clinical parameters between patients with and without ESCC might remain, although we enrolled patients to achieve a uniform age distribution and a similar proportion of alcohol consumption. A prospective study will be required to determine which gene-mutant clones predict future ESCC development and to what extent they contribute to that development. It should be noted that the coverage across gene bodies was not uniform in several genes, and this should be improved to detect mutations more accurately. A germline risk variant of *ALDH2* has been observed exclusively in Japanese and Chinese populations (29), which may render some findings in this study specific to these populations.

Early detection and prevention of cancer are crucial to improving patient outcomes. Our results suggest that non-invasive sampling of the buccal mucosa to measure mucosal remodeling might help identify individuals at high risk for ESCC, thereby enabling targeted surveillance and early intervention.

MATERIALS AND METHODS

Study design

The objectives of this study were to establish somatic mosaicism in the buccal mucosa as a marker reflecting lifestyle and germline risk factors of ESCC and to demonstrate its utility for assessing the presence or absence of ESCC. The individuals studied were patients with and without ESCC who underwent endoscopic examinations. This is an observational case-control study in which the extent of lifestyle habits was matched between the two groups. The investigators were not blinded to the allocation during experiments and outcome assessment

since this is a case-control study. Grades of LVL were independently evaluated by three experienced endoscopists blinded to the sequencing results.

Because this was an exploratory study, no statistical methods were used to determine sample sizes. We enrolled as many patients as possible who provided consent for our study during the enrollment period between December 2019 and April 2024. Sequencing quality was used for inclusion/exclusion criteria, where five samples with a consensus depth $<1,800\times$ were excluded. We did not attempt replication in sequencing experiments but conducted ddPCR experiments for validation-

Patient samples

We first enrolled 62 patients with ESCC or a history of ESCC and 48 subjects without ESCC, confirmed by endoscopic examination within 12 months of sampling, at Kyoto University Hospital and Kobe Asahi Hospital between December 2019 and December 2022 (primary cohort) (table S2). Because patients with ESCC have a higher frequency of heavy drinking, we included both patients with high alcohol consumption and also those with relatively low alcohol consumption. Individuals without ESCC were enrolled to achieve a uniform distribution of age (≥ 40 years old) and a similar proportion of heavy drinking. An additional 112 patients, including 59 with ESCC and 53 without ESCC, were enrolled between January 2023 and April 2024 for model validation (additional cohort) (table S2). In total, 222 patients, including 121 with ESCC and 101 without ESCC were analyzed (fig. S1A and table S2). To investigate the concordance between tissue remodeling in the buccal mucosa and esophagus, we recruited subjects with ESCC ($n=5$) and without ($n=5$) (table S6). Informed consent was obtained from all participants according to the procedures approved by the Internal Review Board of Kyoto University (G0645).

Buccal mucosa samples were collected using three swab sizes with areas of 20, 28, and 254 mm² (fig. S1, B and C). Each swab was rotated 10 times in the same area of the central region of the buccal mucosa to collect the samples. Esophageal mucosa samples were collected through endoscopic biopsy. After removal of the submucosal layer, samples were trimmed to a size of 10 mm². White blood cells were also collected ~~as controls~~ from 20 patients to serve as reference data for filtering out artifacts when calling somatic mutations in buccal swab samples. All blood tests were conducted within 6 months of sampling, except for two cases (9 months and 20 months). Images of esophageal Lugol's staining were obtained during endoscopic examination of all patients with ESCC, except for three patients, and of 15 individuals without ESCC. Grades of Lugol voiding lesion (LVL) were independently evaluated by three experienced endoscopists blinded to the sequencing results, based on the criteria defined previously (31) (grade A: no LVL, grade B: 1-9 LVLs per endoscopic view, grade C: ≥ 10 LVLs). Average ethanol consumption ≥ 18 units/week (1 unit = 22 grams ethanol) was defined as a high-risk drinking habit (heavy drinking), and a smoking intensity ≥ 30 pack-years was defined as a high-risk smoking habit (heavy smoking), where one pack-year is equivalent to one pack per day for one year (31). For patients with ESCC, the stage was determined according to the WHO Classification of Tumors of the Digestive System (5th edition). Stages 0 and 1 ESCC were categorized as early-stage ESCC, while stages 2-4 were classified as advanced ESCC. The characteristics of enrolled patients are shown in tables S1 and S2.

Error-corrected sequencing

We selected 26 genes that were frequently mutated ($\geq 2\%$) and significantly mutated through MutSig CV ($q < 0.05$) or dndscv ($q < 0.1$) in 519 ESCC or 157 physiologically normal esophageal samples analyzed in our previous study (7) (table S8). We also analyzed *ALDH2* and *ADH1B*.

DNA was extracted using the QIAGEN DNA Micro Kit (fig. S1D), followed by library preparation using 25-50 ng of DNA. After fragmentation with the E220evolution Focused-ultrasonicator (Covaris), strand-specific molecular indexing was performed by ligating adapters with double-stranded 8 bp UMIs to independently tag the top and bottom strands of insert DNA (xGen™ Prism DNA Library Prep Kit, IDT). Indexed libraries were hybridized and captured using xGen™ Predesigned Gene Capture Pools (IDT) designed for the 28 genes. Sequencing was conducted on the DNBSEQ-G400RS (MGI Tech) in the 150-bp paired-end mode. The target raw depth of sequencing was 20,000×, and the median raw depth was 45,151× (ranging from 7,665-108,021) (fig. S1E). Sequencing reads were aligned with the human reference genome (GRCh37) using BWA-MEM (v.0.7.17). Read families comprising the top and bottom strands derived from the same original molecules were determined based on the start-stop position and unique sets of UMIs of the reads using fgbio (v.1.4.0) and picard (v.2.24.0) (duplex sequencing (46)). Then, consensus reads were generated while correcting for artefacts, only for read families with ≥ 2 top reads and ≥ 1 bottom read; otherwise, they were discarded. The scripts used to construct consensus reads were deposited in Zenodo (doi: 10.5281/zenodo.14272026). Samples with a consensus read depth $\geq 1,800\times$ were used for this study ($n=707$ from 222 subjects) (fig. S1A). Because mutation detection sensitivity depends on the depth, we randomly downsampled reads to a depth of 2,500× for samples with a depth $\geq 3,000\times$, resulting in a median consensus read depth of 2,497× (ranging from 1,806-2,996) (table S9 and fig. S1E).

The Genomon pipeline (v.2.6.3) (<http://github.com/Genomon-Project>) was used to call somatic mutations, with modifications (<https://github.com/kwatanabe-ku>) made to EBFILTER, to precisely handle high-depth data (v.0.2.5), and to GenomonMutationFilter, to assess artefacts related to cruciform DNA formation (v.0.4.0). To determine the filtering parameters, we

sequenced 12 ESCC samples and 16 independent normal blood samples in the same manner as the buccal swab samples. Somatic mutations with variant allele frequencies (VAFs) ≥ 0.1 detected in the 12 ESCC samples served as gold standards. We mixed data from ESCC samples with normal sample data at ratios ranging from 0.2-1% to simulate mutations with low VAFs. Mutations were called on the virtual data, and filtering parameters were determined to improve sensitivity and reduce the false-positive rate. The final parameters were as follows: (i) VAF ≥ 0.0016 and VAF ≤ 0.3 , (ii) maximum variant read number detected in 20 controls generated from blood DNA ≤ 2 , (iii) ratio of reads realigned to artefactual sequence inferred from cruciform DNA formation < 0.5 , (iii) number of reads realigned to original sequence ≥ 2 , and (iv) significantly higher VAFs than expected for errors ($p \leq 10^{-2.5}$), determined using the EBCall algorithm on the basis of an empirical VAF distribution from 20 controls. This filtering process yielded a sensitivity of 0.957 and a true-positive rate of 0.960 (fig. S1F). After mutation calling for buccal mucosa samples, we further manually inspected mutation loci to remove artifacts. Among variants identified in ≥ 6 samples ($n=616$), we excluded those with $\geq 70\%$ of variant bases located within 50 bases from either end of the reads ($n=44$). In addition, variants with supporting reads ≥ 8 and strand ratio of 0 or 1, where variant reads across mutated samples were considered, were also considered as artifacts and excluded ($n=80$). Because reads with multiple variants are considered artifacts, variants were excluded if the ratio of supporting reads with multiple variants to all supporting reads was ≥ 0.5 or if the average mapping quality of supporting reads was < 44 ($n=854$).

We also performed a validation analysis using digital droplet PCR (ddPCR) with a QX200 (Bio-Rad). We selected *TP53* mutations with validated probes available from the manufacturer. 124 mutations were randomly selected. ddPCR analysis was performed using 50

ng of sample DNA, as well as three control DNA samples. The threshold to discriminate mutant allele signals from wild-type allele signals was determined based on the signal distribution obtained from control samples, and validation criteria were as follows: (i) number of mutant signals in swab sample ≥ 3 ; (ii) number of mutant signals in three controls ≤ 1 ; and (iii) VAF in swab sample $\geq 5 \times$ VAF in three controls. This resulted in a true positive rate of 93.4% (table S4).

The mutant clone size was calculated by multiplying the swab area by $2 \times$ VAF for mutations on autosomes and the X chromosome (female subjects) and by VAF for mutations on the X chromosome (male subjects), assuming no copy number alterations.

Mutational signature analysis

We first employed NanoSeq to unbiasedly detect mutations in buccal swab samples following the protocol (32) with modifications. In brief, DNA digested with HpyCH4V underwent A-tailing with ddBTPs and was ligated to adapters containing UMIs using the Twist cfDNA Library Preparation Kit (Twist). After quantifying the libraries by qPCR, sequencing was conducted on the DNBSEQ-G400RS in the 150-bp paired-end mode. The data was processed with the NanoSeq pipeline (v.3.5.4) (<https://github.com/cancerit/NanoSeq>), using default settings except for the UMI length, which was changed from 3 to 5 bp. We randomly selected 16 samples previously analyzed by duplex sequencing from subjects with various drinking and smoking habits, both with and without germline risks, followed by NanoSeq analysis (table S5). In total, 15,578 single nucleotide variants (median 827 (624-2,043)) were identified. De novo mutational signatures were extracted using SigProfilerExtractor (v.1.1.23) and were further decomposed into the COSMIC SBS signatures (v3.4). Then, somatic mutations detected in this study were decomposed into the extracted SBS signatures using the R package MutationalPatterns

(v.3.10.0). Solitary samples were utilized for comparisons between germline and lifestyle risk groups, while all samples were used for LME model analyses.

Whole-exome sequencing (WES)

Samples were collected using 0.2 mm² punch biopsy (BP-A05F, KAI) from the buccal mucosa, followed by DNA extraction using the DNA Micro Kit. Subsequently, a matched blood sample was collected. After enzymatically fragmenting 1.2-10 ng of DNA, libraries were prepared with xGenTM Lotus DNA Library Prep Kit (IDT). Indexed libraries were hybridized and captured using xGenTM Exome Hyb Panel v2 (IDT). Sequencing was conducted on a DNBSEQ-G400RS in 150 bp paired-end mode. The target depth was 100×, and the average actual depth was 85× (21–139×) (table S10). Mutation calling was performed using the Genomon pipeline, as previously described (7). Filtering parameters used in Genomon were as follows: (i) mapping quality ≥ 20 ; (ii) base call quality ≥ 15 ; (iv) total reads ≥ 10 and variant reads ≥ 3 ; (v) VAFs ≥ 0.05 (for tissue sample) and ≤ 0.02 (for normal control); (vi) strand ratio $\neq 0$ or 1; (vii) EBcall $p \leq 10^{-4}$ using WES data of non-paired peripheral blood samples ($n=16$); and (viii) Fisher's test $p \leq 10^{-1}$. To identify expanded clones across multiple samples, each of the mutations detected in a specimen was interrogated in all the remaining samples; if a variant detected in one or more samples by Genomon was also present (VAF ≥ 0.05) in another sample within the biopsy specimen, the variant was considered to be true-positive. The mutant cell fraction was calculated by multiplying the sampling area by $2 \times \text{VAF}$ for mutations on autosomes and the X chromosome (female subjects) and by VAF for mutations on the X chromosome (male subjects) because there were no copy number alterations assessed by CNACS (7).

External datasets

Somatic mutations identified in 519 ESCC samples and 157 physiologically normal esophageal samples were obtained from our previous study (7). The MutSig CV and dndscv results were also obtained.

Germline risk loci of *ALDH2* and *ADH1B*

ALDH2 rs671 and *ADH1B* rs1229984 are prevalent functional variants in East Asian populations. The risk and protective genotypes were defined as rs671 (AG, risk; AA and GG, protective) and rs1229984 (GG, risk; AA and AG, protective), respectively (47). *ALDH2* rs671(AG) and *ADH1B* rs1229984 (GG) were defined as high-risk germline variants.

Model construction for assessing the presence of ESCC

We initially conducted univariate logistic regression analyses on solitary samples in the primary cohort for each swab type to examine whether the number or sum of VAFs of driver mutations in each positively selected gene could predict the presence of ESCC. Significant variables ($p < 0.05$) were selected for further analysis. After 75% of the samples were randomly chosen, multivariable logistic regression analysis was conducted. The optimal combination of variables was then determined using the LASSO method with the R package glmnet (v.4.1.8). A set of stable variables selected in more than 80% of ten thousand trials was used to construct models. Internal cross-validation was conducted to evaluate the validity of the models by randomly splitting the entire sample into training (75%) and validation (25%) sets ten thousand times. Test models were constructed using the training sets, and c-indices were calculated for the validation sets. For validation, propensity score matching was employed to select pairs with similar characteristics, including age, sex, status of ESCC and HNSCC, alcohol consumption, smoking intensity, MCV, and germline variants of *ALDH2* and *ADH1B*, from the primary and additional

cohorts before model construction, aiming to reduce selection bias [MatchIt (v.4.5.5)]. Models were constructed using matched data from the primary cohort and subsequently applied to the additional cohort data. Optimal thresholds were defined as the points closest to (0, 1) on the ROC curves for the primary cohort. The R code used to construct the models was deposited in Zenodo (doi: 10.5281/zenodo.14272026).

Statistical analyses

Statistical analyses were performed using R (v.4.2.3). All P-values were calculated using two-sided analysis unless otherwise specified. The Mann—Whitney *U* test or Chi-square test were used for group comparisons. LVL grade comparison was performed, excluding missing values (complete case analysis). To examine the uniformity of buccal mucosal remodeling, we applied linear regression to compare numbers and sums of VAFs of mutations from the left and right buccal mucosa. dN/dS values were calculated using the R package dndscv (v.0.0.1.0), where q-values were obtained by a two-sided likelihood ratio test and multiple testing correction based on the Benjamini—Hochberg method. All data were included in the analyses unless otherwise specified. Outliers were defined as data points exceeding 1.5 times the interquartile range above the third quartile or below the first quartile. LME models and generalized LME models were assessed using the R package lmerTest (v.3.1.3), where solitary samples were used for analyses with swab sizes as random effects. The P-values for both linear regression and logistic regression were calculated using a *t*-test and a Wald test, respectively, to assess whether each regression coefficient significantly differs from zero. Model summaries are shown in tables S11 and S12.

List of Supplementary Materials

Figs. S1 to S9

Tables S1 to S12

MDAR Reproducibility Checklist

References and Notes

1. N. Kakiuchi, S. Ogawa, Clonal expansion in non-cancer tissues. *Nat Rev Cancer* **21**, 239-256 (2021).
2. H. Maeda, N. Kakiuchi, Clonal expansion in normal tissues. *Cancer Sci* **115**, 2117-2124 (2024).
3. L. C. Fernandez, M. Torres, F. X. Real, Somatic mosaicism: on the road to cancer. *Nat Rev Cancer* **16**, 43-55 (2016).
4. G. Genovese, A. K. Kahler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landen, M. Hoglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. Gronberg, C. M. Hultman, S. A. McCarroll, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).
5. S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C. H. Mermel, N. Burt, A. Chavez, J. M. Higgins, V. Moltchanov, F. C. Kuo, M. J. Kluk, B. Henderson, L. Kinnunen, H. A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B. F. Banahan, S. Gabriel, S. Kathiresan, H. M. Stringham, M. I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman, L. Groop, G. Atzmon, J. G. Wilson, D. Neuberg, D.

- Altshuler, B. L. Ebert, Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498 (2014).
6. I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M. R. Stratton, P. H. Jones, P. J. Campbell, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
 7. A. Yokoyama, N. Kakiuchi, T. Yoshizato, Y. Nannya, H. Suzuki, Y. Takeuchi, Y. Shiozawa, Y. Sato, K. Aoki, S. K. Kim, Y. Fujii, K. Yoshida, K. Kataoka, M. M. Nakagawa, Y. Inoue, T. Hirano, Y. Shiraishi, K. Chiba, H. Tanaka, M. Sanada, Y. Nishikawa, Y. Amanuma, S. Ohashi, I. Aoyama, T. Horimatsu, S. Miyamoto, S. Tsunoda, Y. Sakai, M. Narahara, J. B. Brown, Y. Sato, G. Sawada, K. Mimori, S. Minamiguchi, H. Haga, H. Seno, S. Miyano, H. Makishima, M. Muto, S. Ogawa, Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312-317 (2019).
 8. A. R. J. Lawson, F. Abascal, T. H. H. Coorens, Y. Hooks, L. O'Neill, C. Latimer, K. Raine, M. A. Sanders, A. Y. Warren, K. T. A. Mahbubani, B. Bareham, T. M. Butler, L. M. R. Harvey, A. Cagan, A. Menzies, L. Moore, A. J. Colquhoun, W. Turner, B. Thomas, V. Gnanapragasam, N. Williams, D. M. Rassl, H. Vohringer, S. Zumalave, J. Nangalia, J. M. C. Tubio, M. Gerstung, K. Saeb-Parsy, M. R. Stratton, P. J. Campbell, T. J. Mitchell, I. Martincorena, Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75-82 (2020).

9. R. Li, Y. Du, Z. Chen, D. Xu, T. Lin, S. Jin, G. Wang, Z. Liu, M. Lu, X. Chen, T. Xu, F. Bai, Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* **370**, 82-89 (2020).
10. N. Kakiuchi, K. Yoshida, M. Uchino, T. Kihara, K. Akaki, Y. Inoue, K. Kawada, S. Nagayama, A. Yokoyama, S. Yamamoto, M. Matsuura, T. Horimatsu, T. Hirano, N. Goto, Y. Takeuchi, Y. Ochi, Y. Shiozawa, Y. Kogure, Y. Watatani, Y. Fujii, S. K. Kim, A. Kon, K. Kataoka, T. Yoshizato, M. M. Nakagawa, A. Yoda, Y. Nanya, H. Makishima, Y. Shiraishi, K. Chiba, H. Tanaka, M. Sanada, E. Sugihara, T. A. Sato, T. Maruyama, H. Miyoshi, M. M. Taketo, J. Oishi, R. Inagaki, Y. Ueda, S. Okamoto, H. Okajima, Y. Sakai, T. Sakurai, H. Haga, S. Hirota, H. Ikeuchi, H. Nakase, H. Marusawa, T. Chiba, O. Takeuchi, S. Miyano, H. Seno, S. Ogawa, Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**, 260-265 (2020).
11. K. Nanki, M. Fujii, M. Shimokawa, M. Matano, S. Nishikori, S. Date, A. Takano, K. Toshimitsu, Y. Ohta, S. Takahashi, S. Sugimoto, K. Ishimaru, K. Kawasaki, Y. Nagai, R. Ishii, K. Yoshida, N. Sasaki, T. Hibi, S. Ishihara, T. Kanai, T. Sato, Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *Nature* **577**, 254-259 (2020).
12. H. Lee-Six, S. Olafsson, P. Ellis, R. J. Osborne, M. A. Sanders, L. Moore, N. Georgakopoulos, F. Torrente, A. Noorani, M. Goddard, P. Robinson, T. H. H. Coorens, L. O'Neill, C. Alder, J. Wang, R. C. Fitzgerald, M. Zilbauer, N. Coleman, K. Saeb-Parsy, I. Martincorena, P. J. Campbell, M. R. Stratton, The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).

13. L. Moore, D. Leongamornlert, T. H. H. Coorens, M. A. Sanders, P. Ellis, S. C. D'Ento, K. J. Dawson, T. Butler, R. Rahbari, T. J. Mitchell, F. Maura, J. Nangalia, P. S. Tarpey, S. F. Brunner, H. Lee-Six, Y. Hooks, S. Moody, K. T. Mahbubani, M. Jimenez-Linan, J. J. Brosens, C. A. Iacobuzio-Donahue, I. Martincorena, K. Saeb-Parsy, P. J. Campbell, M. R. Stratton, The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646 (2020).
14. M. Yamaguchi, H. Nakaoka, K. Suda, K. Yoshihara, T. Ishiguro, N. Yachida, K. Saito, H. Ueda, K. Sugino, Y. Mori, K. Yamawaki, R. Tamura, S. Revathi, T. Motoyama, K. Tainaka, R. G. W. Verhaak, I. Inoue, T. Enomoto, Spatiotemporal dynamics of clonal selection and diversification in normal endometrial epithelium. *Nat Commun* **13**, 943 (2022).
15. K. Yoshida, K. H. C. Gowers, H. Lee-Six, D. P. Chandrasekharan, T. Coorens, E. F. Maughan, K. Beal, A. Menzies, F. R. Millar, E. Anderson, S. E. Clarke, A. Pennycuik, R. M. Thakrar, C. R. Butler, N. Kakiuchi, T. Hirano, R. E. Hynds, M. R. Stratton, I. Martincorena, S. M. Janes, P. J. Campbell, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272 (2020).
16. T. Nishimura, N. Kakiuchi, K. Yoshida, T. Sakurai, T. R. Kataoka, E. Kondoh, Y. Chigusa, M. Kawai, M. Sawada, T. Inoue, Y. Takeuchi, H. Maeda, S. Baba, Y. Shiozawa, R. Saiki, M. M. Nakagawa, Y. Nannya, Y. Ochi, T. Hirano, T. Nakagawa, Y. Inagaki-Kawata, K. Aoki, M. Hirata, K. Nanki, M. Matano, M. Saito, E. Suzuki, M. Takada, M. Kawashima, K. Kawaguchi, K. Chiba, Y. Shiraishi, J. Takita, S. Miyano, M. Mandai, T. Sato, K. Takeuchi, H. Haga, M. Toi, S. Ogawa, Evolutionary histories of breast cancer and related clones. *Nature* **620**, 607-614 (2023).

17. S. F. Brunner, N. D. Roberts, L. A. Wylie, L. Moore, S. J. Aitken, S. E. Davies, M. A. Sanders, P. Ellis, C. Alder, Y. Hooks, F. Abascal, M. R. Stratton, I. Martincorena, M. Hoare, P. J. Campbell, Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538-542 (2019).
18. A. G. Bick, J. S. Weinstock, S. K. Nandakumar, C. P. Fulco, E. L. Bao, S. M. Zekavat, M. D. Szeto, X. Liao, M. J. Leventhal, J. Nasser, K. Chang, C. Laurie, B. B. Burugula, C. J. Gibson, A. E. Lin, M. A. Taub, F. Aguet, K. Ardlie, B. D. Mitchell, K. C. Barnes, A. Moscati, M. Fornage, S. Redline, B. M. Psaty, E. K. Silverman, S. T. Weiss, N. D. Palmer, R. S. Vasan, E. G. Burchard, S. L. R. Kardia, J. He, R. C. Kaplan, N. L. Smith, D. K. Arnett, D. A. Schwartz, A. Correa, M. de Andrade, X. Guo, B. A. Konkle, B. Custer, J. M. Peralta, H. Gui, D. A. Meyers, S. T. McGarvey, I. Y. Chen, M. B. Shoemaker, P. A. Peyser, J. G. Broome, S. M. Gogarten, F. F. Wang, Q. Wong, M. E. Montasser, M. Daya, E. E. Kenny, K. E. North, L. J. Launer, B. E. Cade, J. C. Bis, M. H. Cho, J. Lasky-Su, D. W. Bowden, L. A. Cupples, A. C. Y. Mak, L. C. Becker, J. A. Smith, T. N. Kelly, S. Aslibekyan, S. R. Heckbert, H. K. Tiwari, I. V. Yang, J. A. Heit, S. A. Lubitz, J. M. Johnsen, J. E. Curran, S. E. Wenzel, D. E. Weeks, D. C. Rao, D. Darbar, J. Y. Moon, R. P. Tracy, E. J. Buth, N. Rafaels, R. J. F. Loos, P. Durda, Y. Liu, L. Hou, J. Lee, P. Kachroo, B. I. Freedman, D. Levy, L. F. Bielak, J. E. Hixson, J. S. Floyd, E. A. Whitsel, P. T. Ellinor, M. R. Irvin, T. E. Fingerlin, L. M. Raffield, S. M. Armasu, M. M. Wheeler, E. C. Sabino, J. Blangero, L. K. Williams, B. D. Levy, W. H. Sheu, D. M. Roden, E. Boerwinkle, J. E. Manson, R. A. Mathias, P. Desai, K. D. Taylor, A. D. Johnson, N. T.-O. f. P. M. Consortium, P. L. Auer, C. Kooperberg, C. C. Laurie, T. W. Blackwell, A. V. Smith, H. Zhao, E. Lange, L. Lange, S. S. Rich, J. I. Rotter, J. G.

- Wilson, P. Scheet, J. O. Kitzman, E. S. Lander, J. M. Engreitz, B. L. Ebert, A. P. Reiner, S. Jaiswal, G. Abecasis, V. G. Sankaran, S. Kathiresan, P. Natarajan, Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763-768 (2020).
19. M. D. Kessler, A. Damask, S. O'Keeffe, N. Banerjee, D. Li, K. Watanabe, A. Marketta, M. Van Meter, S. Semrau, J. Horowitz, J. Tang, J. A. Kosmicki, V. M. Rajagopal, Y. Zou, Y. Houvras, A. Ghosh, C. Gillies, J. Mbatchou, R. R. White, N. Verweij, J. Bovijn, N. N. Parikshak, M. G. LeBlanc, M. Jones, C. Regeneron Genetics, G.-R. D. Collaboration, D. J. Glass, L. A. Lotta, M. N. Cantor, G. S. Atwal, A. E. Locke, M. A. R. Ferreira, R. Deering, C. Paulding, A. R. Shuldiner, G. Thurston, A. A. Ferrando, W. Salerno, J. G. Reid, J. D. Overton, J. Marchini, H. M. Kang, A. Baras, G. R. Abecasis, E. Jorgenson, Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301-309 (2022).
 20. J. Pegliasco, P. Hirsch, C. Marzac, F. Isnard, J. C. Meniane, C. Deswarte, P. Pellet, C. Lemaitre, G. Leroy, G. Rabadan Moraes, H. Guermouche, B. Schmaltz-Panneau, F. Pasquier, C. Colas, P. R. Benusiglio, O. Bera, J. H. Bourhis, E. Brissot, O. Caron, S. Chraïbi, P. Cony-Makhoul, C. Delaunay-Darivon, S. Lapusan, F. S. de Fontbrune, P. Fuseau, A. Najman, W. Vainchenker, F. Delhommeau, J. B. Micol, I. Plo, C. Bellanne-Chantelot, Germline ATG2B/GSKIP-containing 14q32 duplication predisposes to early clonal hematopoiesis leading to myeloid neoplasms. *Leukemia* **36**, 126-137 (2022).
 21. I. Martincorena, J. C. Fowler, A. Wabik, A. R. J. Lawson, F. Abascal, M. W. J. Hall, A. Cagan, K. Murai, K. Mahbubani, M. R. Stratton, R. C. Fitzgerald, P. A. Handford, P. J. Campbell, K. Saeb-Parsy, P. H. Jones, Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917 (2018).

22. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021).
23. E. Morgan, I. Soerjomataram, H. Rungay, H. G. Coleman, A. P. Thrift, J. Vignat, M. Laversanne, J. Ferlay, M. Arnold, The Global Landscape of Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma Incidence and Mortality in 2020 and Projections to 2040: New Estimates From GLOBOCAN 2020. *Gastroenterology* **163**, 649-658 e642 (2022).
24. N. Gupta, A. Bansal, S. B. Wani, S. Gaddam, A. Rastogi, P. Sharma, Endoscopy for upper GI cancer screening in the general population: a cost-utility analysis. *Gastrointest Endosc* **74**, 610-624 e612 (2011).
25. P. C. Enzinger, R. J. Mayer, Esophageal cancer. *N Engl J Med* **349**, 2241-2252 (2003).
26. M. Matejcic, M. J. Gunter, P. Ferrari, Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. *Carcinogenesis* **38**, 859-872 (2017).
27. S. Ishiguro, S. Sasazuki, M. Inoue, N. Kurahashi, M. Iwasaki, S. Tsugane, J. S. Group, Effect of alcohol consumption, cigarette smoking and flushing response on esophageal cancer risk: a population-based cohort study (JPHC study). *Cancer Lett* **275**, 240-246 (2009).
28. D. P. Slaughter, H. W. Southwick, W. Smejkal, Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* **6**, 963-968 (1953).

29. S. Ohashi, S. Miyamoto, O. Kikuchi, T. Goto, Y. Amanuma, M. Muto, Recent Advances From Basic and Clinical Studies of Esophageal Squamous Cell Carcinoma. *Gastroenterology* **149**, 1700-1715 (2015).
30. Y. Lu, T. Sobue, T. Kitamura, R. Matsuse, Y. Kitamura, K. Matsuo, H. Ito, I. Oze, T. Shimazu, T. Yamaji, M. Iwasaki, S. Sasazuki, N. Sawada, S. Tsugane, Cigarette smoking, alcohol drinking, and oral cavity and pharyngeal cancer in the Japanese: a population-based cohort study in Japan. *Eur J Cancer Prev* **27**, 171-179 (2018).
31. C. Katada, T. Yokoyama, T. Yano, K. Kaneko, I. Oda, Y. Shimizu, H. Doyama, T. Koike, K. Takizawa, M. Hirao, H. Okada, T. Yoshii, K. Konishi, T. Yamanouchi, T. Tsuda, T. Omori, N. Kobayashi, T. Shimoda, A. Ochiai, Y. Amanuma, S. Ohashi, T. Matsuda, H. Ishikawa, A. Yokoyama, M. Muto, Alcohol Consumption and Multiple Dysplastic Lesions Increase Risk of Squamous Cell Carcinoma in the Esophagus, Head, and Neck. *Gastroenterology* **151**, 860-869 e867 (2016).
32. F. Abascal, L. M. R. Harvey, E. Mitchell, A. R. J. Lawson, S. V. Lensing, P. Ellis, A. J. C. Russell, R. E. Alcantara, A. Baez-Ortega, Y. Wang, E. J. Kwa, H. Lee-Six, A. Cagan, T. H. H. Coorens, M. S. Chapman, S. Olafsson, S. Leonard, D. Jones, H. E. Machado, M. Davies, N. F. Obro, K. T. Mahubani, K. Allinson, M. Gerstung, K. Saeb-Parsy, D. G. Kent, E. Laurenti, M. R. Stratton, R. Rahbari, P. J. Campbell, R. J. Osborne, I. Martincorena, Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405-410 (2021).
33. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganello, R. Sabarinathan, D. A. Wheeler,

- V. Mustonen, P. M. S. W. Group, G. Getz, S. G. Rozen, M. R. Stratton, P. Consortium, Author Correction: The repertoire of mutational signatures in human cancer. *Nature* **614**, E41 (2023).
34. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganello, R. Sabarinathan, D. A. Wheeler, V. Mustonen, P. M. S. W. Group, G. Getz, S. G. Rozen, M. R. Stratton, P. Consortium, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
 35. J. Chang, W. Tan, Z. Ling, R. Xi, M. Shao, M. Chen, Y. Luo, Y. Zhao, Y. Liu, X. Huang, Y. Xia, J. Hu, J. S. Parker, D. Marron, Q. Cui, L. Peng, J. Chu, H. Li, Z. Du, Y. Han, W. Tan, Z. Liu, Q. Zhan, Y. Li, W. Mao, C. Wu, D. Lin, Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat Commun* **8**, 15290 (2017).
 36. S. Moody, S. Senkin, S. M. A. Islam, J. Wang, D. Nasrollahzadeh, R. Cortez Cardoso Penha, S. Fitzgerald, E. N. Bergstrom, J. Atkins, Y. He, A. Khandekar, K. Smith-Byrne, C. Carreira, V. Gaborieau, C. Latimer, E. Thomas, I. Abnizova, P. E. Bucciarelli, D. Jones, J. W. Teague, B. Abedi-Ardekani, S. Serra, J. Y. Scoazec, H. Saffar, F. Azmoudeh-Ardalan, M. Sotoudeh, A. Nikmanesh, H. Poustchi, A. Niavarani, S. Gharavi, M. Eden, P. Richman, L. S. Campos, R. C. Fitzgerald, L. F. Ribeiro, S. C. Soares-Lima, C. Dzamalala, B. T. Mmbaga, T. Shibata, D. Menya, A. M. Goldstein, N. Hu, R. Malekzadeh, A. Fazel, V. McCormack, J. McKay, S. Perdomo, G. Scelo, E. Chanudet, L. Humphreys, L. B. Alexandrov, P. Brennan, M. R. Stratton, Mutational signatures in

- esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat Genet* **53**, 1553-1563 (2021).
37. I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, P. J. Campbell, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021 (2017).
 38. C. Katada, T. Yokoyama, T. Yano, I. Oda, Y. Shimizu, H. Doyama, T. Koike, K. Takizawa, M. Hirao, H. Okada, T. Yoshii, Y. Kubota, T. Yamanouchi, T. Tsuda, T. Omori, N. Kobayashi, H. Suzuki, S. Tanabe, K. Hori, N. Nakayama, H. Kawakubo, N. Kakushima, Y. Matsuo, H. Ishikawa, A. Yokoyama, M. Muto, Association between macrocytosis and metachronous squamous cell carcinoma of the esophagus after endoscopic resection in men with early esophageal squamous cell carcinoma. *Esophagus* **17**, 149-158 (2020).
 39. J. C. Fowler, C. King, C. Bryant, M. W. J. Hall, R. Sood, S. H. Ong, E. Earp, D. Fernandez-Antoran, J. Koeppel, S. C. Dentre, D. Shorthouse, A. Durrani, K. Fife, E. Rytina, D. Milne, A. Roshan, K. Mahububani, K. Saeb-Parsy, B. A. Hall, M. Gerstung, P. H. Jones, Selection of Oncogenic Mutant Clones in Normal Human Skin Varies with Body Site. *Cancer Discov* **11**, 340-361 (2021).
 40. Y. Usui, Y. Taniyama, M. Endo, Y. N. Koyanagi, Y. Kasugai, I. Oze, H. Ito, I. Imoto, T. Tanaka, M. Tajika, Y. Niwa, Y. Iwasaki, T. Aoi, N. Hakozaiki, S. Takata, K. Suzuki, C. Terao, M. Hatakeyama, M. Hirata, K. Sugano, T. Yoshida, Y. Kamatani, H. Nakagawa, K. Matsuda, Y. Murakami, A. B. Spurdle, K. Matsuo, Y. Momozawa, Helicobacter pylori, Homologous-Recombination Genes, and Gastric Cancer. *N Engl J Med* **388**, 1181-1190 (2023).

41. P. Kleiblova, I. A. Shaltiel, J. Benada, J. Sevcik, S. Pechackova, P. Pohlreich, E. E. Voest, P. Dundr, J. Bartek, Z. Kleibl, R. H. Medema, L. Macurek, Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. *J Cell Biol* **201**, 511-521 (2013).
42. S. L. Harris, A. J. Levine, The p53 pathway: positive and negative feedback loops. *Oncogene* **24**, 2899-2908 (2005).
43. A. Yokoyama, C. Katada, T. Yokoyama, T. Yano, K. Kaneko, I. Oda, Y. Shimizu, H. Doyama, T. Koike, K. Takizawa, M. Hirao, H. Okada, T. Yoshii, K. Konishi, T. Yamanouchi, T. Tsuda, T. Omori, N. Kobayashi, H. Suzuki, S. Tanabe, K. Hori, N. Nakayama, H. Kawakubo, H. Ishikawa, M. Muto, Alcohol abstinence and risk assessment for second esophageal cancer in Japanese men after mucosectomy for early esophageal cancer. *Plos One* **12**, (2017).
44. A. Yokoyama, C. Katada, T. Yokoyama, K. Takizawa, T. Yano, I. Oda, Y. Shimizu, H. Nakanishi, T. Koike, M. Hirao, H. Okada, T. Yoshii, A. Katagiri, T. Yamanouchi, Y. Matsuo, H. Kawakubo, N. Kobayashi, H. Ishikawa, M. Muto, The Alcohol Use Disorders Identification Test and the risk of metachronous cancer after endoscopic resection of esophageal cancer. *Carcinogenesis* **41**, 1049-1056 (2020).
45. O. Bugter, M. C. W. Spaander, M. J. Bruno, R. J. Baatenburg de Jong, A. Amelink, D. J. Robinson, Optical detection of field cancerization in the buccal mucosa of patients with esophageal cancer. *Clin Transl Gastroenterol* **9**, 152 (2018).
46. S. R. Kennedy, M. W. Schmitt, E. J. Fox, B. F. Kohn, J. J. Salk, E. H. Ahn, M. J. Prindle, K. J. Kuong, J. C. Shen, R. A. Risques, L. A. Loeb, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606 (2014).

47. R. Cui, Y. Kamatani, A. Takahashi, M. Usami, N. Hosono, T. Kawaguchi, T. Tsunoda, N. Kamatani, M. Kubo, Y. Nakamura, K. Matsuda, Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology* **137**, 1768-1775 (2009).

Acknowledgments: We thank T. Ida, A. Ryu, T. Shirahari, C. Sakurai, F. Pu, and K. Koishihara for technical assistance; E. Shichino and all staff in the outpatient units at Kobe Asahi Hospital for their support in sample collection.

Funding:

Japan Agency for Medical Research and Development (AMED) JP23ck0106798 (NK), JP24ama221434 (AY), and JP24zf0127009 (SO)

Japan Society for the Promotion of Science (JSPS) KAKENHI JP20H03513 and JP22K19457 (NK), JP20H03660 (AY), and JP19H05656 and JP24H00009 (SO)

Japan Science and Technology Agency (JST): Fusion Oriented REsearch for disruptive Science and Technology (FOREST) Program JPMJFR215V (NK) and Moonshot Research and Development Program JPMJMS2022-25 (NK) Takeda Science Foundation (AY, NK)

Princess Takamatsu Cancer Research Fund (AY, NK)

Terumo Life Science Foundation (AY)

Inoue Foundation for Science (NK)

The Naito Foundation (NK)

Japanese Cancer Association – Kobayashi Foundation for Cancer Research (NK)

Nanpuh Hospital (SO)

Author contributions:

Conceptualization: AY, NK

Methodology: AY, KW, YInoue, SK, YN, NK

Software: KW, NK

Validation: AY, NK

Formal analysis: AY, NK

Investigation: AY, KW, TH, YIshida, YInoue, YT, NK

Resources: AY, MT, KH, YK, SKK, CK

Data curation: AY, NK

Writing – original draft: AY, NK

Writing – review & editing: AY, NK

Visualization: AY, NK

Supervision: AY, HS, SO, MM, NK

Project administration: AY, NK

Funding acquisition: AY, NK

Competing interests: This work is protected under a patent application currently pending approval (Application Number: JP2024-061000, title: Method and Kit for Predicting the

Onset and/or Assessing the Risk of Esophageal Cancer, Pharyngeal Cancer, or Oral Cancer, applicant: Kyoto University, coinventors: AY, NK, SO, MM). The other authors declare that they have no competing interests.

Data and materials availability: All data associated with this study are in the paper or supplementary materials. Bam files have been deposited in the European Genome-Phenome Archive (EGA) under accession numbers EGAS500000000054, EGAS500000000055, EGAS500000000443, EGAS500000000717, and EGAS500000000718. To access these data, users require approval from the Data Access Committee for sequencing data at Kyoto University (<https://ega-archive.org/dacs/EGAC500000000045>). The users can apply for the use of data via the above URL, where the policy for granting data is also described. The R codes used to construct the models and scripts used to construct consensus reads were deposited in Zenodo (doi: 10.5281/zenodo.14272026).

Figure legends

Fig. 1. Somatic mutations in esophageal driver genes in normal buccal mucosa. (A)

Explanation of study design. (B) Number (top) and relative frequency (bottom) of mutations allocated to SBS1, SBS5/SBS40a, and SBS16 are drawn for solitary samples of the primary cohort, stratified by swab size. Information about germline risk variants of *ALDH2* and *ADH1B*, high-risk alcohol consumption, high-risk smoking intensity, ESCC, HNSCC, LVL grade C, and MCV ≥ 106 fL is indicated in the middle. (C) The effect of risk factors on the number of mutations was assessed using a linear mixed-effects model. The p-value for each risk factor is plotted (*t*-test), with the fixed effects and 95% CI shown on the right. (D and E) Number of mutations (D) and SBS16 mutations (E) are plotted against alcohol consumption (ethanol unit/week), stratified by the presence (red) or absence (blue) of germline risk variants. The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model are indicated for each risk group.

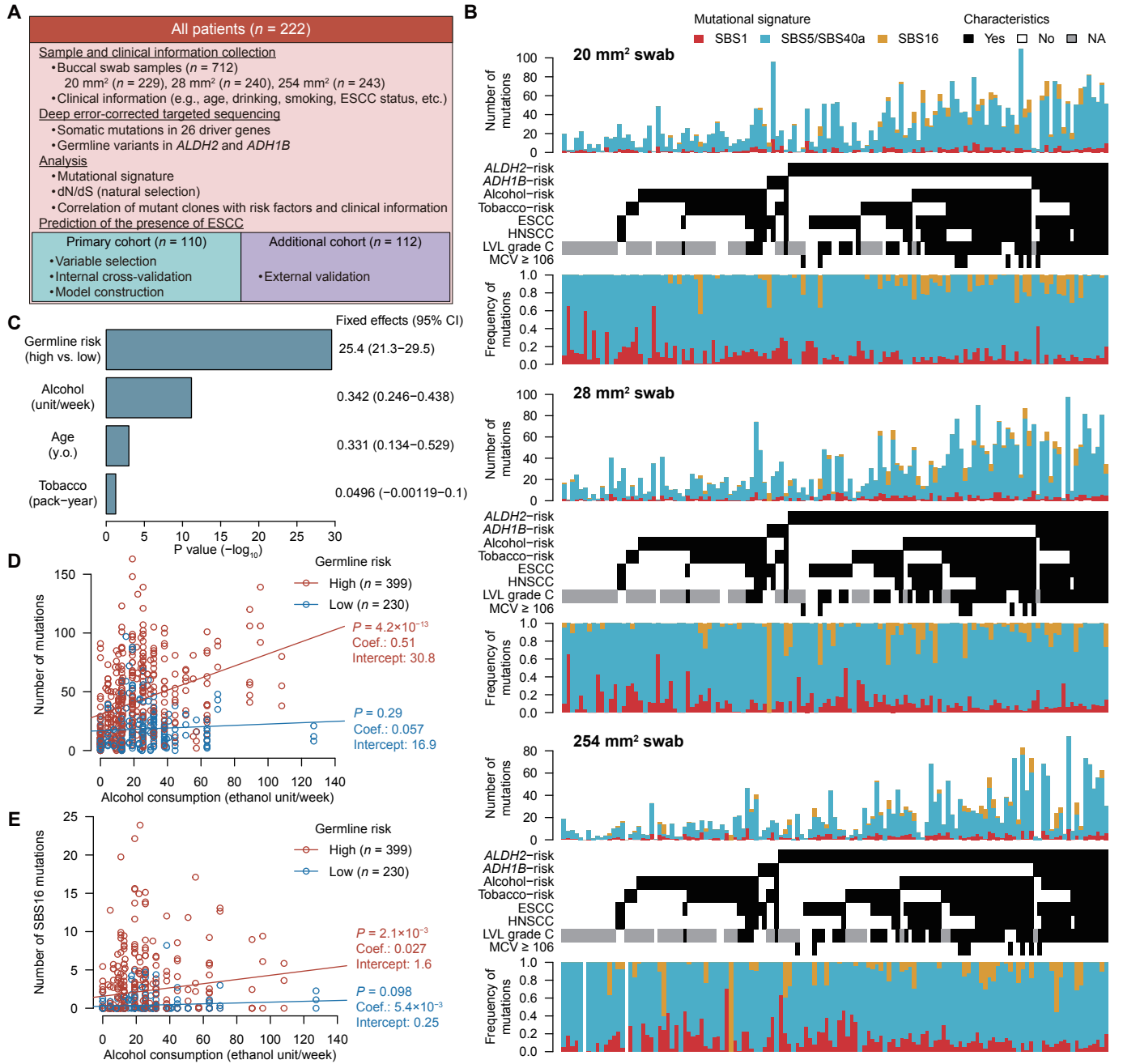
Fig. 2. Positively selected genes in buccal mucosa. (A) Number of mutations detected in each gene among all samples, with mutation types indicated by color. (B) dN/dS values for each gene and mutation type are plotted. Asterisks indicate dN/dS > 1 with $q < 0.05$, calculated by a two-sided likelihood ratio test and multiple testing correction based on the Benjamini—Hochberg method using *dndscv* in R. (C) Distribution of mutations in 707 buccal mucosa (top), 157 physiologically normal esophagus (PNE) (middle), and 519 ESCC (bottom) samples for the positively selected genes *TP53*, *NOTCH1*, *CHEK2*, and *PPM1D*. Colors and shapes of the lollipops indicate mutation types. Average consensus depths of each amino acid in duplex sequencing are shown for buccal mucosa (light blue bars). Domains within genes are indicated by colors.

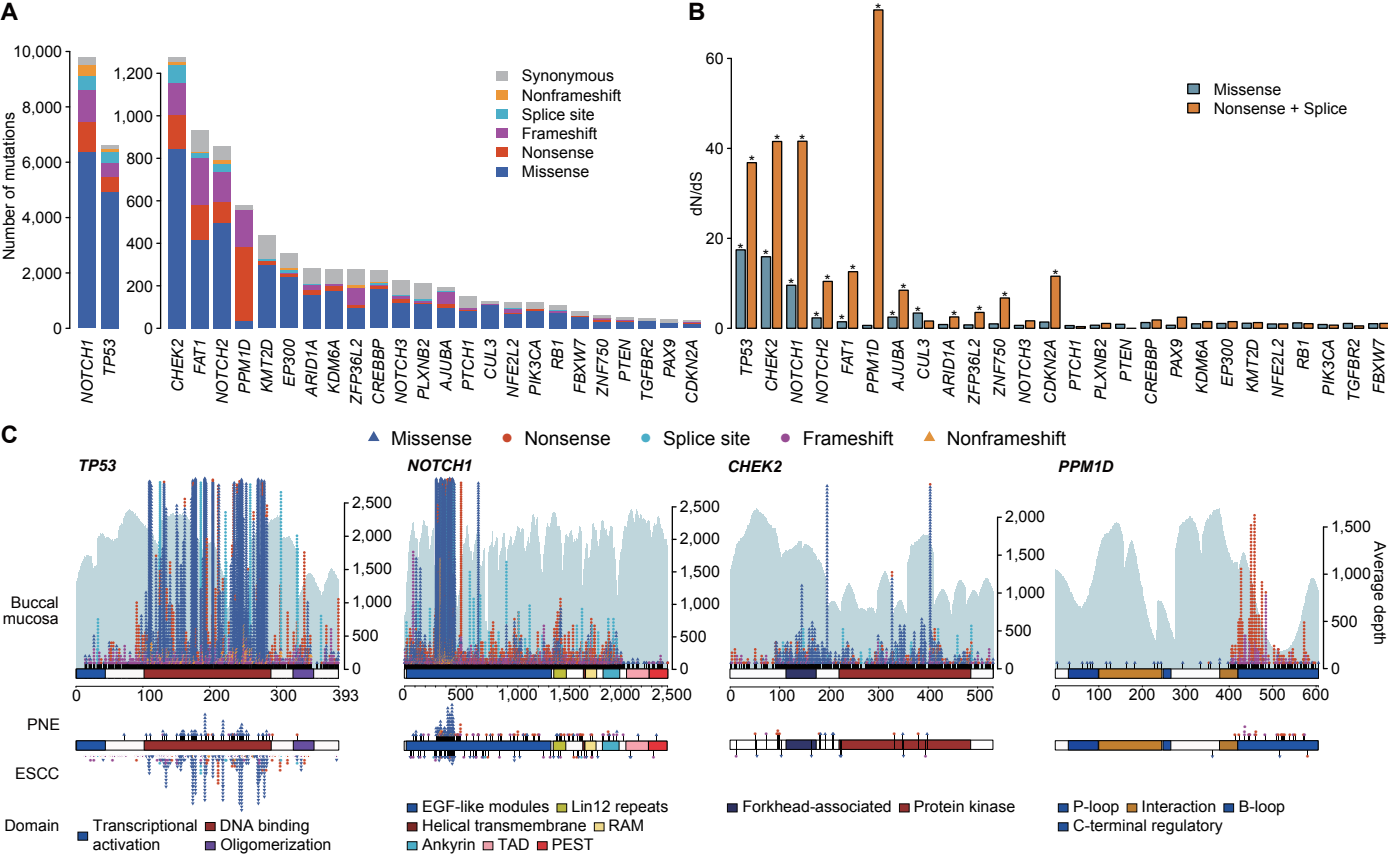
Fig. 3. Correlation between driver mutations and ESCC risk factors. (A) The correlation of driver mutations in positively selected genes with known risk factors for ESCC using a linear mixed-effects model in terms of the number of driver mutations (shown in orange) and sum of their VAFs (shown in green) (*t*-test). (B) Number (top) and sum of VAFs (bottom) of driver mutations in *TP53* plotted against alcohol consumption (ethanol unit/week) in samples from individuals with low- (blue) and high-risk (red) variants of *ALDH2*. The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model are indicated. (C) The correlation of driver mutations in positively selected genes with age and lifestyle risk factors for ESCC using a linear mixed-effects model in terms of the number of driver mutations (orange) and sum of their VAFs (green) (*t*-test), stratified by germline risks.

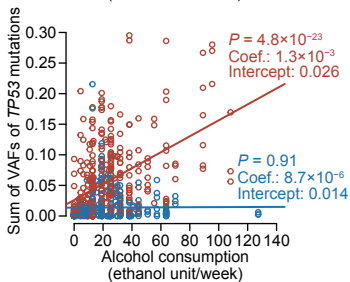
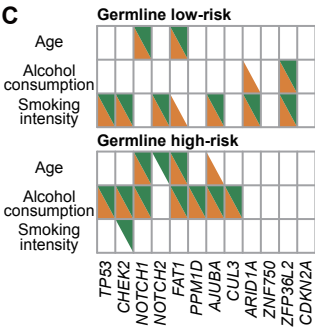
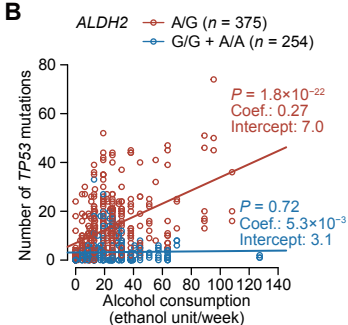
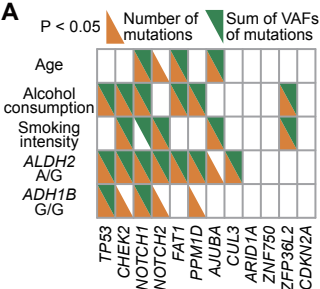
Fig. 4. Correlation of buccal mucosal remodeling with lifestyle habits and clinical characteristics related to ESCC. (A and B) Number (A) and sum of VAFs (B) of driver mutations plotted against alcohol consumption (left) and smoking intensity (middle) for samples from subjects with germline low- (blue) and high-risk variants (red). The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model for each risk group are indicated. Estimates of coefficients of ethanol consumption, smoking intensity, and age in the models are also shown (right), with 95% CI and p-values (*t*-test). (C) Histograms of number (left) and sum of VAFs (right) of driver mutations in solitary samples, with colors distinguishing the presence and absence of ESCC and germline risk variants. (D) Number (left) and sum of VAFs (right) of driver mutations are plotted for individuals without ESCC ($n = 88, 97, \text{ and } 98$ for 20, 28, and 254 mm² swabs, respectively) and those with early ($n = 84, 86, \text{ and } 88$ for 20, 28, and 254 mm² swabs,

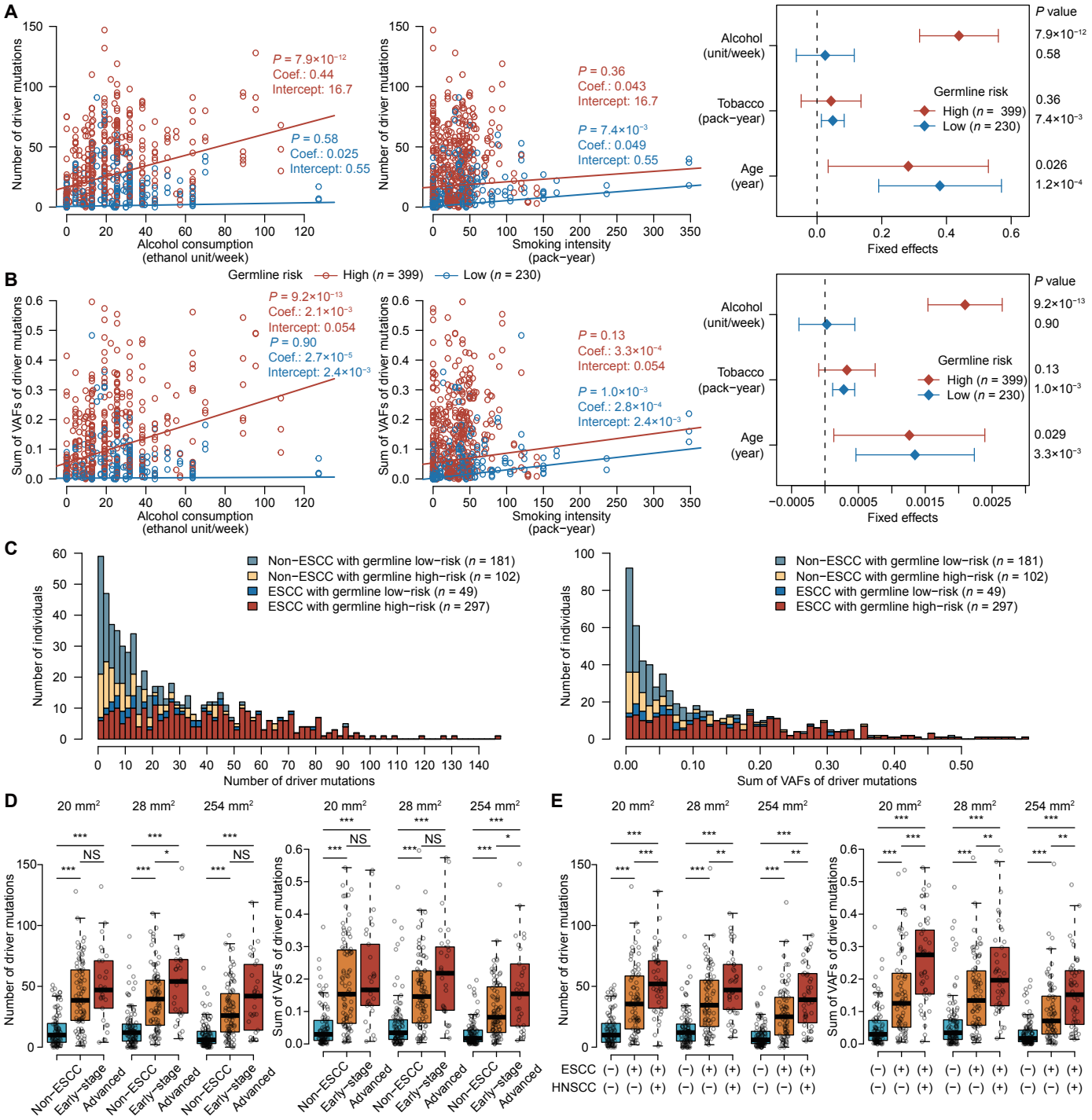
respectively) or advanced ESCC ($n = 29, 29,$ and 30 for $20, 28,$ and 254 mm^2 swabs, respectively) stratified by swab size. **(E)** Number (left) and sum of VAFs (right) plotted for individuals without ESCC ($n = 88, 97,$ and 98 for $20, 28,$ and 254 mm^2 swabs, respectively) and those with only ESCC ($n = 72, 74,$ and 74 for $20, 28,$ and 254 mm^2 swabs, respectively), or both ESCC and HNSCC ($n = 41, 41,$ and 44 for $20, 28,$ and 254 mm^2 swabs, respectively). Box plots represent the median, first quartile, and third quartile, with whiskers extending to the furthest value within a $1.5\times$ interquartile range. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and NS=not significant, by Mann–Whitney U test.

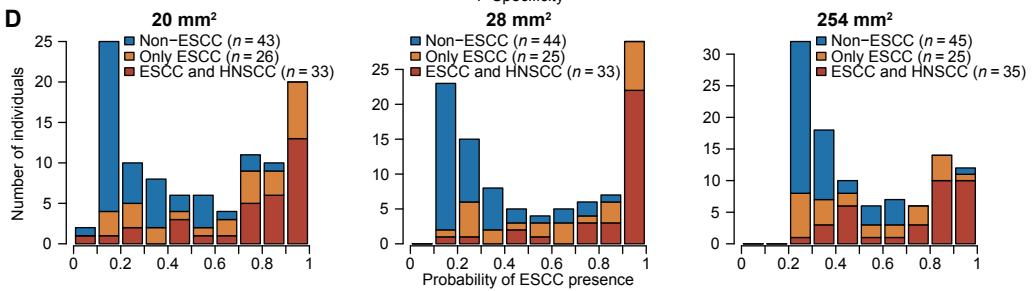
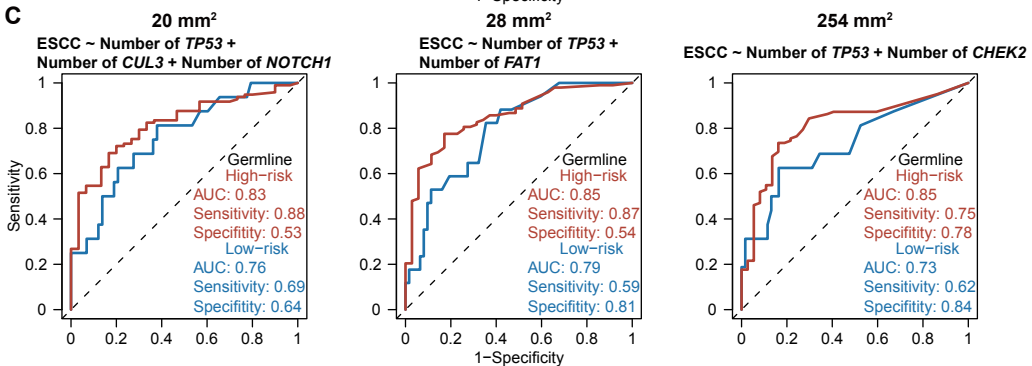
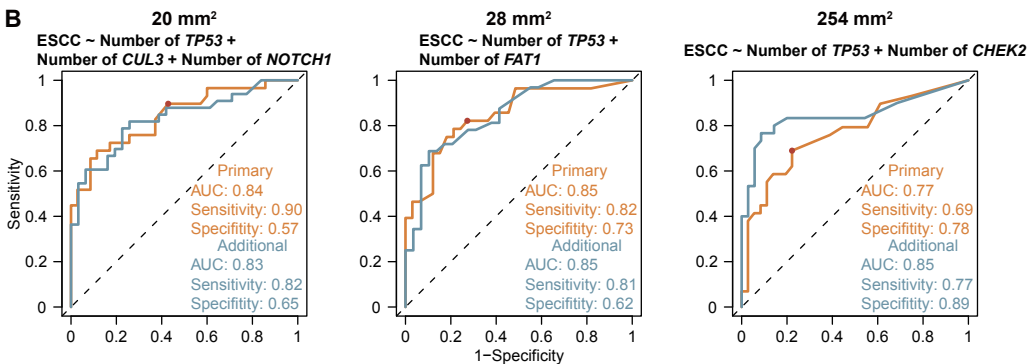
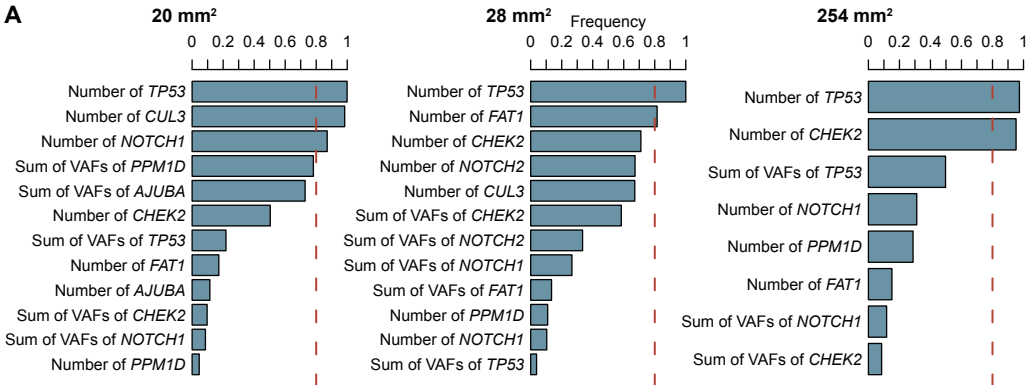
Fig. 5. Models to assess the presence of ESCC based on driver-mutant clones in the buccal mucosa. **(A)** The fraction of trials in which variables were selected in the multivariable logistic regression model using LASSO for each swab type. The red dotted line was drawn at a frequency of 0.8. **(B)** For each swab type, ROC curves for the primary (orange) and the additional (blue) cohorts were drawn using models constructed based on the matched data from the primary cohort. Red points indicate thresholds determined as the points closest to (0, 1) on the ROC curves for the primary cohort. The AUC, sensitivity, and specificity at the threshold are indicated. **(C)** For each swab type, ROC curves for individuals with (red) and without (blue) germline risk variants were drawn using the same models as (B). The AUC, sensitivity, and specificity at the same threshold are indicated. **(D)** Histograms of the probability of ESCC presence in the primary cohort, calculated by the models, are plotted for each swab type, with colors distinguishing individuals without ESCC, those with only ESCC, and those with both ESCC and HNSCC.











Supplementary Materials for

Somatic mosaicism in the buccal mucosa reflects lifestyle and germline risk factors for esophageal squamous cell carcinoma

Akira Yokoyama, Koichi Watanabe, Yoshikage Inoue, Tomonori Hirano, Masashi Tamaoki, Kenshiro Hirohashi, Shun Kawaguchi, Yoshihiro Ishida, Yasuhide Takeuchi, Yo Kishimoto, Soo Ki Kim, Chikatoshi Katada, Yasuhito Nannya, Hiroshi Seno, Seishi Ogawa, Manabu Muto, Nobuyuki Kakiuchi

Corresponding authors: Akira Yokoyama, yokoaki@kuhp.kyoto-u.ac.jp; Nobuyuki Kakiuchi, kakiuchi@kuhp.kyoto-u.ac.jp

List of Supplementary Materials:

Figs. S1 to S9

Other files associated with these Supplementary Materials:

Tables S1 to S12

MDAR reproducibility checklist

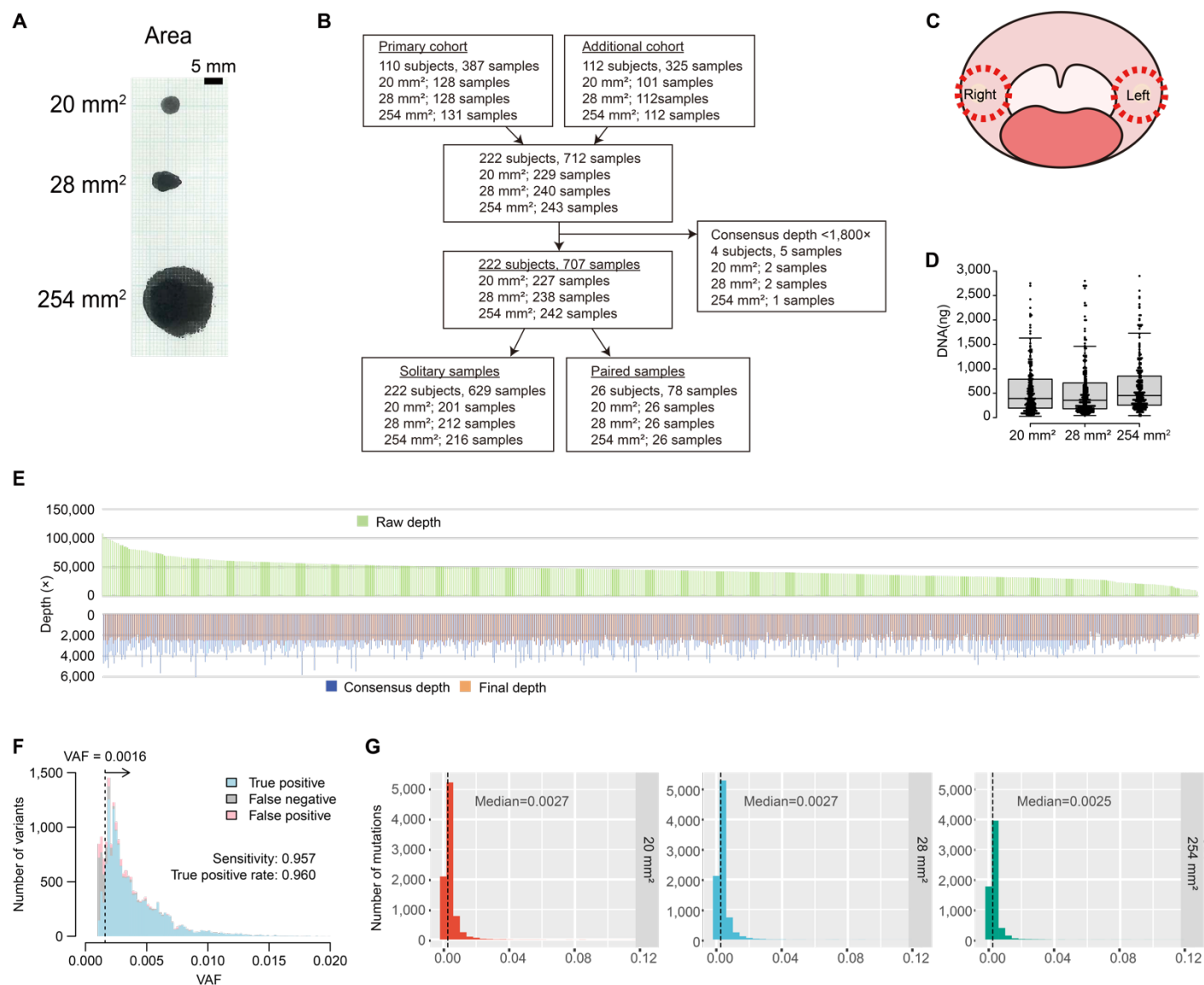
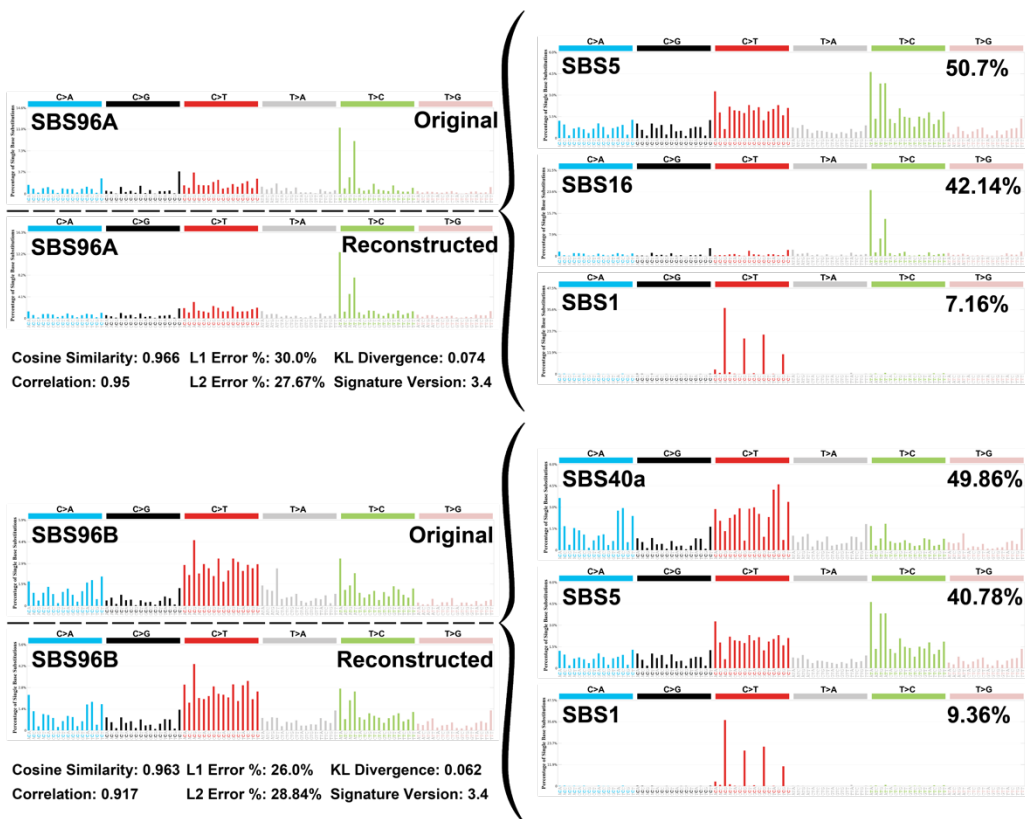
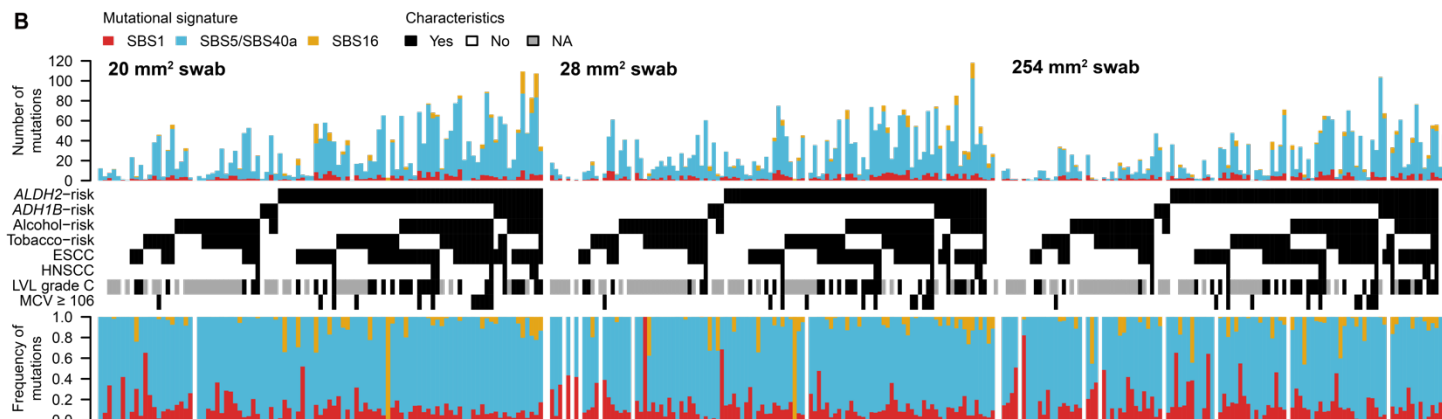


Fig. S1. Samples and sequencing results. (A) Areas of three different swab sizes. Scale bar = 5mm. (B) In total, 712 samples were obtained from 222 individuals (110 in the primary cohort and 112 in the additional cohort) using swabs of three different sizes. After excluding samples with a low consensus depth (<1,800×) of error-corrected sequencing, 707 samples were used for further analyses. Among these, 78 paired samples from 26 individuals were obtained from the opposite side of the buccal mucosa. (C) Illustration of the sampling sites. (D) Amounts of DNA obtained from samples collected using 20, 28, and 254 mm² swabs. (E) Bar plots of raw depths (top), consensus depths (bottom, blue), and final depths (bottom, orange) after downsampling for samples with consensus depth $\geq 3,000\times$ of error-corrected sequencing. (F) Histogram of true positive, false negative, and false positive variants of simulated data using optimized filtering settings (see Materials and Methods). Sensitivity and true positive rate for mutations with VAF ≥ 0.016 are indicated. (G) Histogram of VAFs of mutations stratified by swab size.

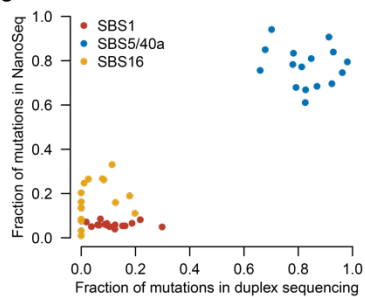
A



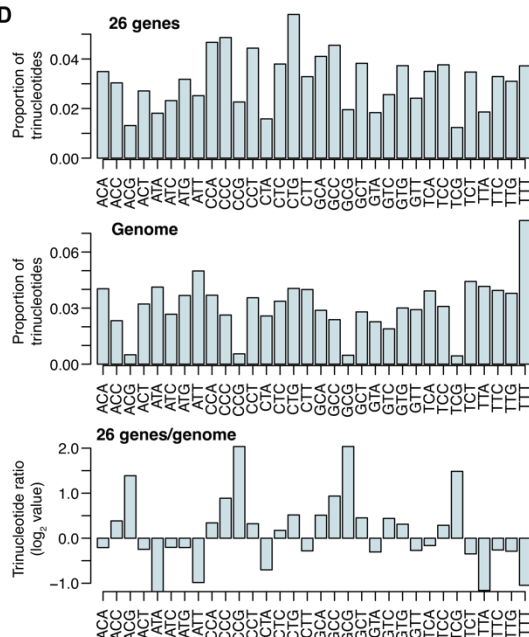
B



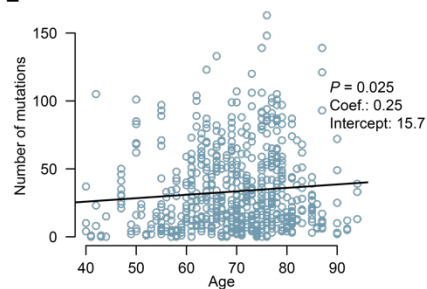
C



D



E



F

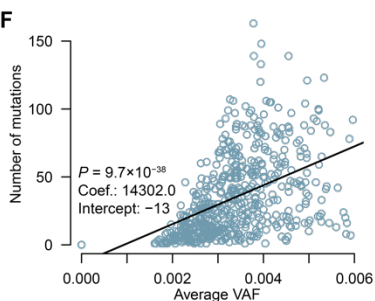


Fig. S2. Mutational signature extraction in normal buccal mucosa. (A) Two *de novo* mutational signatures extracted using SigProfilerExtractor, which were further decomposed into the COSMIC SBS signatures. (B) Number (top) and relative frequency (bottom) of mutations allocated to SBS1, SBS5/SBS40a, and SBS16 are drawn for samples from the additional cohort, stratified by swab size. Information about germline risk variants of *ALDH2* and *ADH1B*, high-risk alcohol consumption, high-risk smoking intensity, ESCC, HNSCC, LVL grade C, and MCV ≥ 106 fL is indicated in the middle. (C) Fractions of mutations attributed to SBS1, SBS5/SBS40a, and SBS16 are plotted for samples analyzed by both duplex sequencing and NanoSeq. (D) Proportions of trinucleotides in the 26 genes analyzed (top) and in the genome (middle), along with trinucleotide ratios of the 26 genes to the genome (bottom). (E) Number of mutations plotted against age (year) in solitary samples. The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model are indicated. (F) Number of mutations plotted against average VAF in solitary samples after excluding outliers with high average VAFs ($n=596$). The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model are indicated.

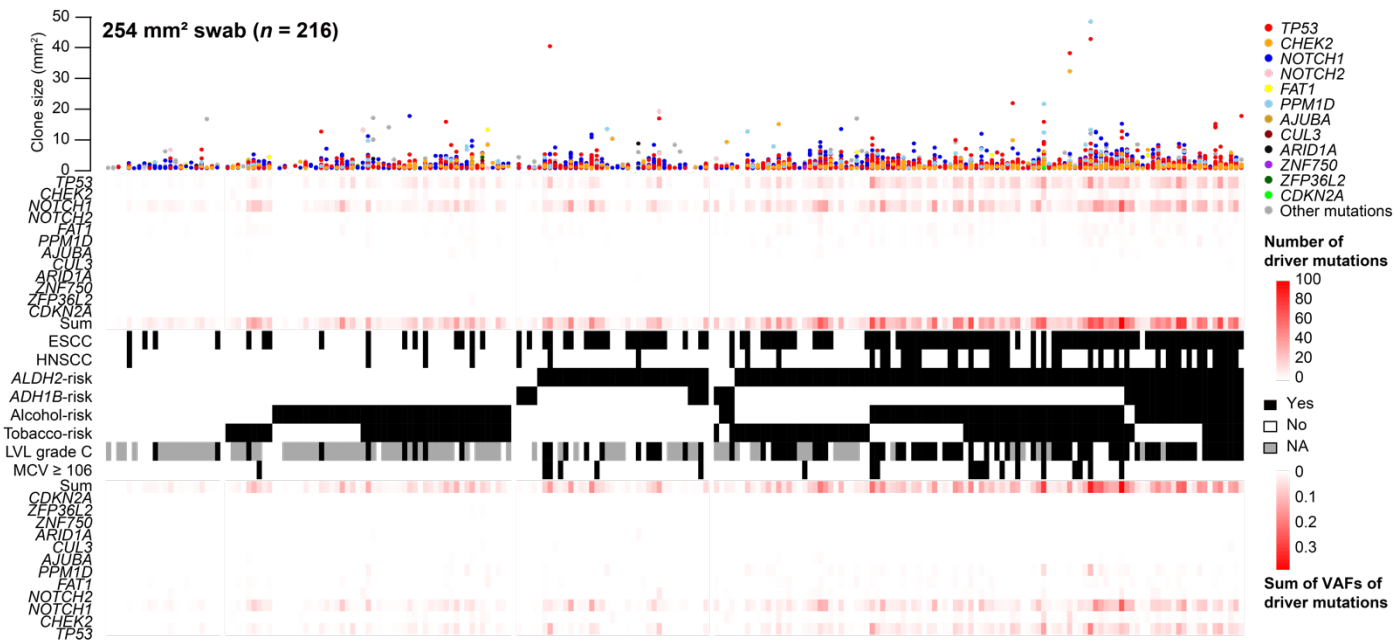
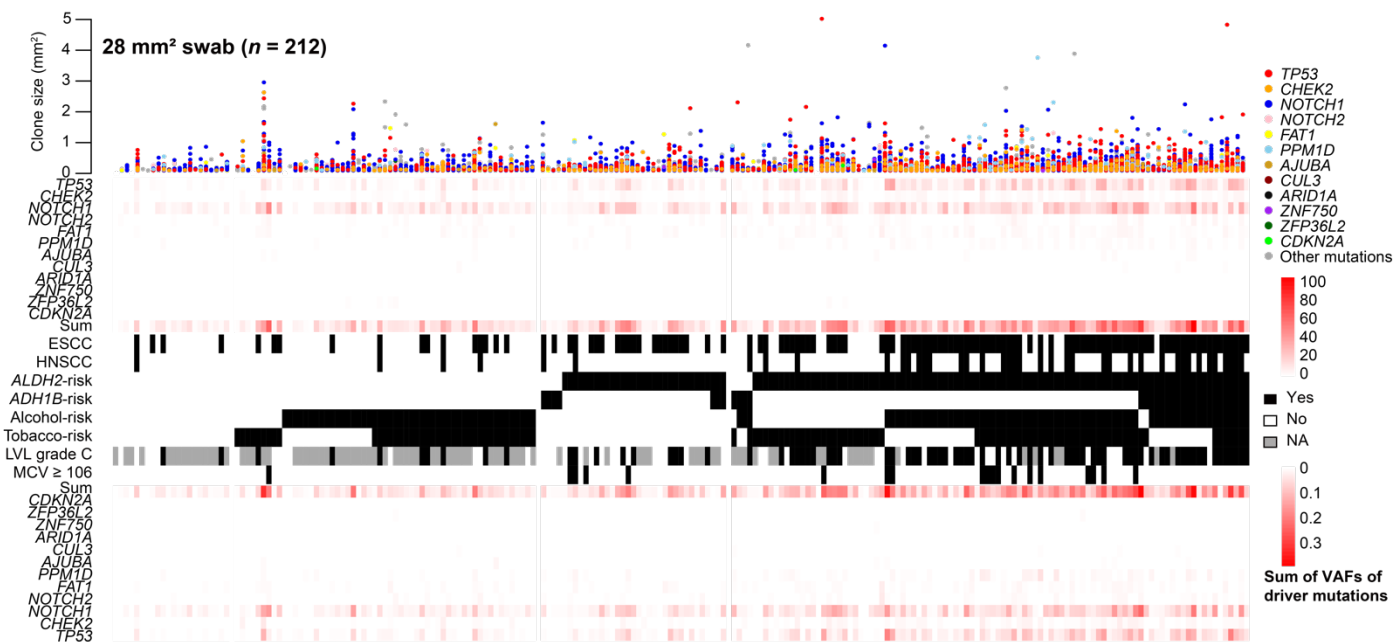
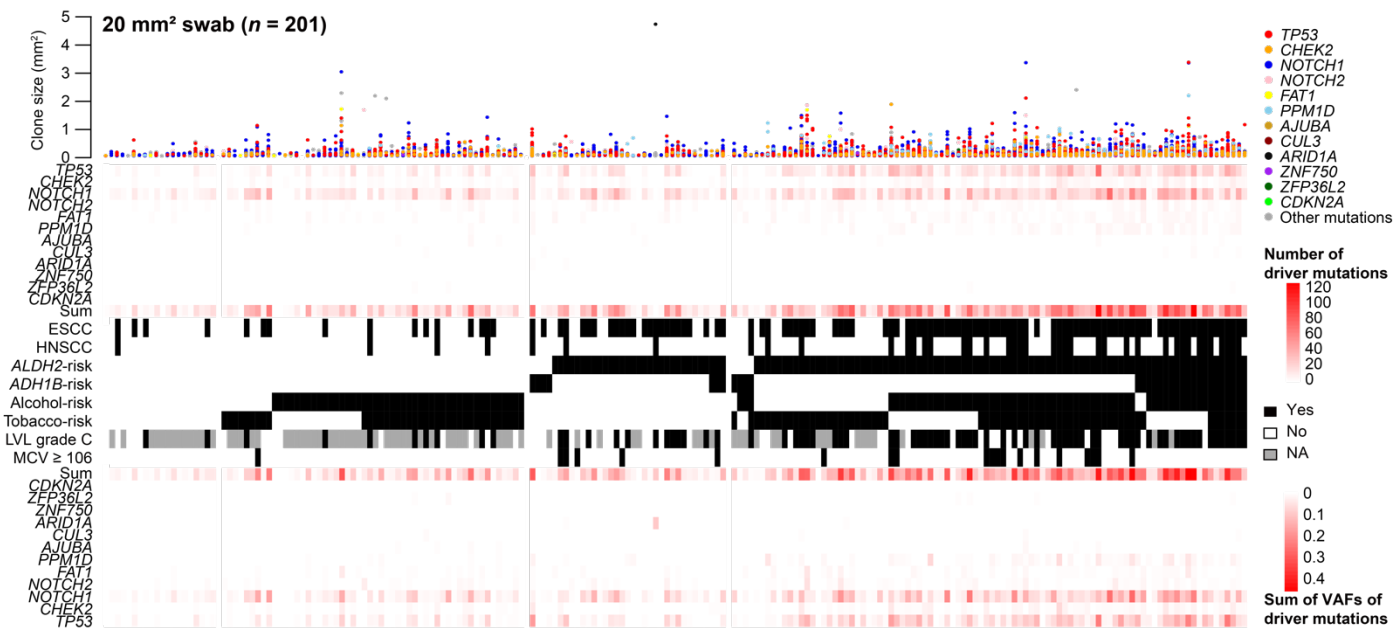


Fig. S3. Summary of driver mutations in normal buccal mucosa. Clone sizes of the mutations are plotted for three types of swabs (top), where colors indicate driver genes. Heatmaps for the number (middle) and sum of VAFs (bottom) of positively selected genes are shown. Information on ESCC, HNSCC, risk variants of *ALDH2* and *ADH1B*, high-risk alcohol consumption, high-risk smoking intensity, MCV ≥ 106 fL, and LVL grade C is also shown.

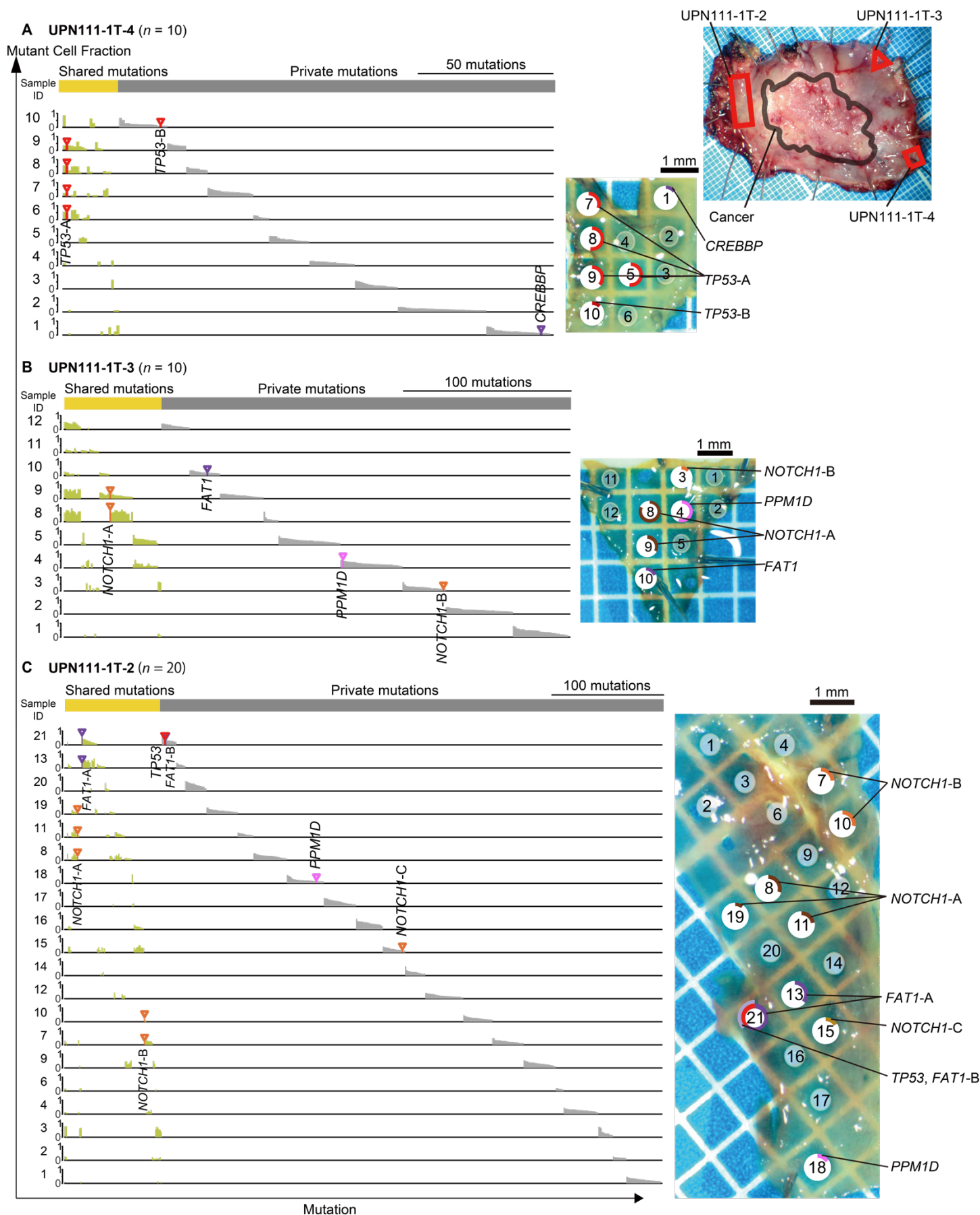


Fig. S4. Clonal expansion in buccal mucosa analyzed through multi-sampling. (A to C) Multiple samples ($n = 10, 10, 20$) were collected from three pieces of normal buccal mucosa of a 58-year-old patient who underwent

buccal mucosal resection because of cancer. Samples were collected in 1-mm lattices using a 0.2 mm² punch biopsy. Mutant cell fractions (MCFs) of mutations detected using WES are plotted (left). The sampling positions are indicated (right). Driver mutations are discriminated by color, and MCFs of driver mutations are indicated by the arch length in the sampling maps. Different mutations in the same gene are discriminated by suffixes (-A, -B, -C) attached to gene symbols. Calibration is provided for each specimen.

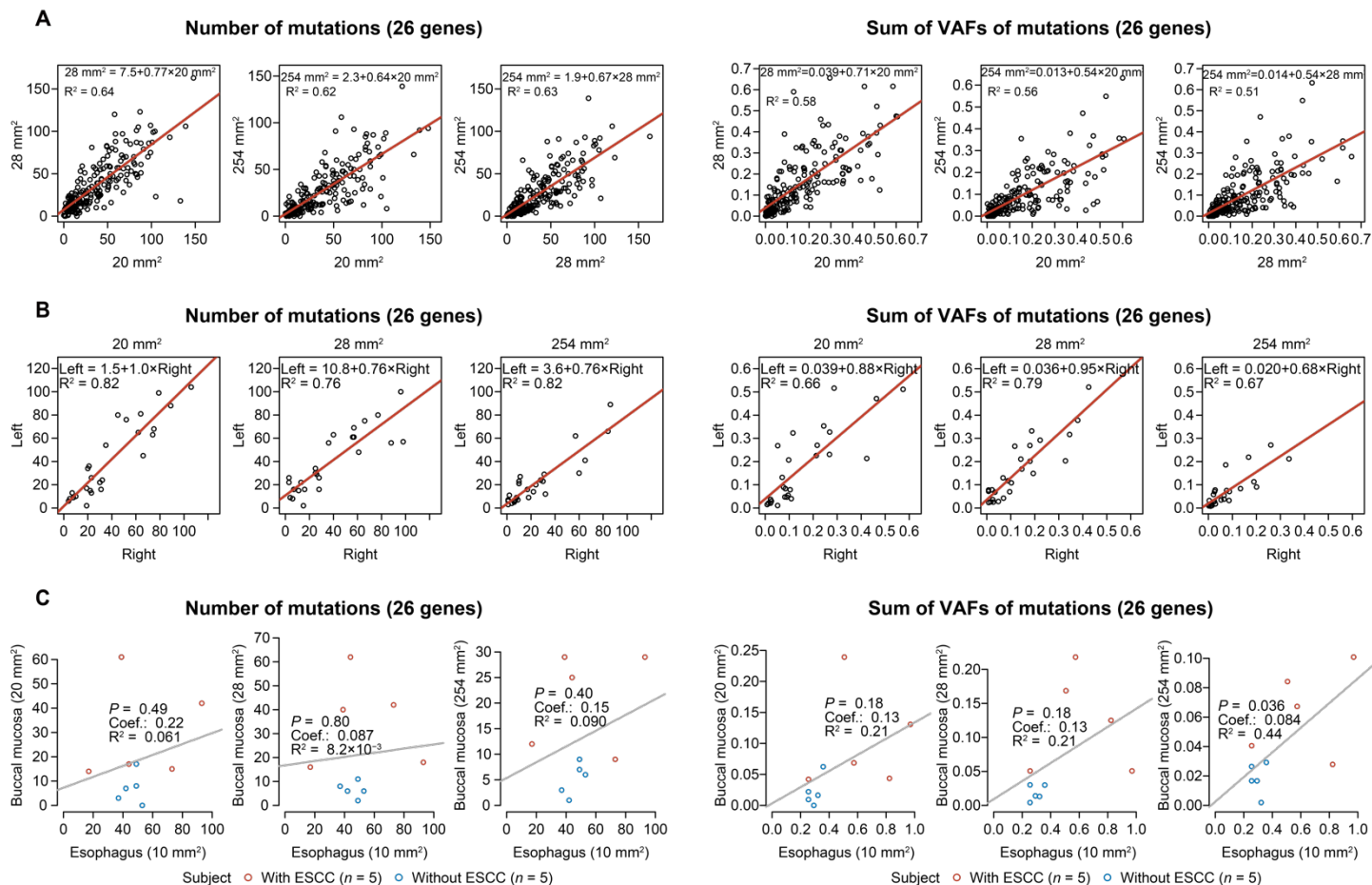


Fig. S5. Uniformity of buccal mucosal remodeling within individuals and comparison of tissue remodeling between buccal and esophageal mucosa. (A) The number (left) and sum of VAFs (right) of mutations in the 26 genes were compared between samples obtained from the same individuals using different swab sizes. The red line represents the regression line. The slope, intercept, and R² value for the regression are indicated. (B) The number (left) and sum of VAFs (right) of mutations in the 26 genes were plotted for the paired samples, stratified by swab size. The red line represents the regression line. The slope, intercept, and R² value for the regression are indicated. (C) The number (left) and sum of VAFs (right) of mutations in the 26 genes were compared between samples obtained from the esophagus and buccal mucosa of the same individuals. The red line represents the regression line. The p-value (*t*-test), coefficient, and R² value for the regression are indicated. Samples obtained from patients with and without ESCC are presented in red and blue, respectively.

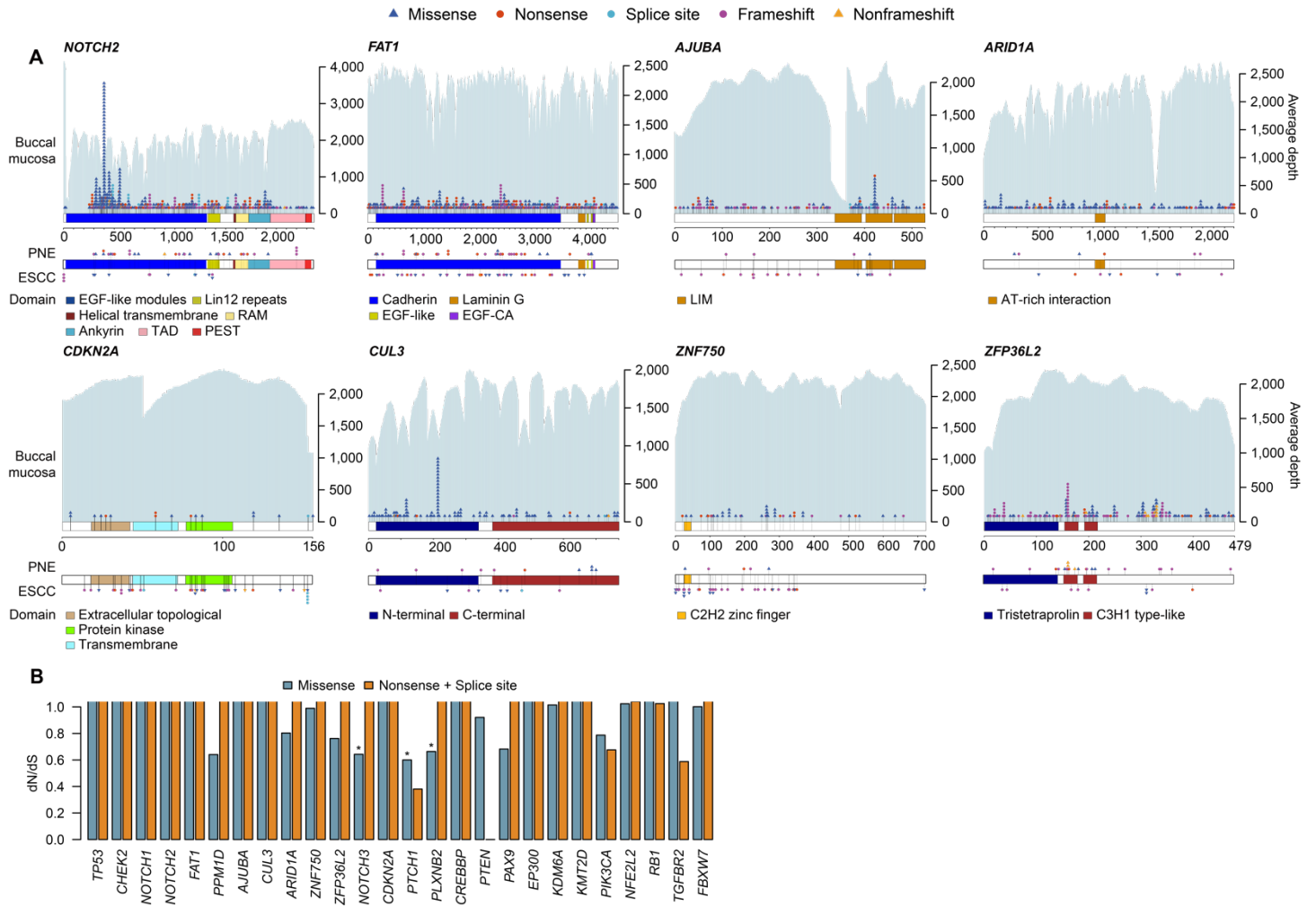


Fig. S6. Mutation distribution and negative selection in buccal mucosa. (A) Distribution of mutations in 707 buccal mucosa (top), 157 physiologically normal esophagus (PNE) (middle), and 519 ESCC (bottom) samples for *NOTCH2*, *FAT1*, *AJUBA*, *ARID1A*, *CDKN2A*, *CUL3*, *ZNF750*, and *ZFP36L2*. Colors and shapes of the lollipops indicate mutation types. Average consensus depths of each amino acid in duplex sequencing are shown for buccal mucosa (light blue bars). Domains within genes are indicated by colors. **(B)** dN/dS values for each gene and mutation type are plotted. Asterisks indicate dN/dS < 1 with $q < 0.05$, calculated by a two-sided likelihood ratio test and multiple testing correction based on the Benjamini–Hochberg method using dndscv.

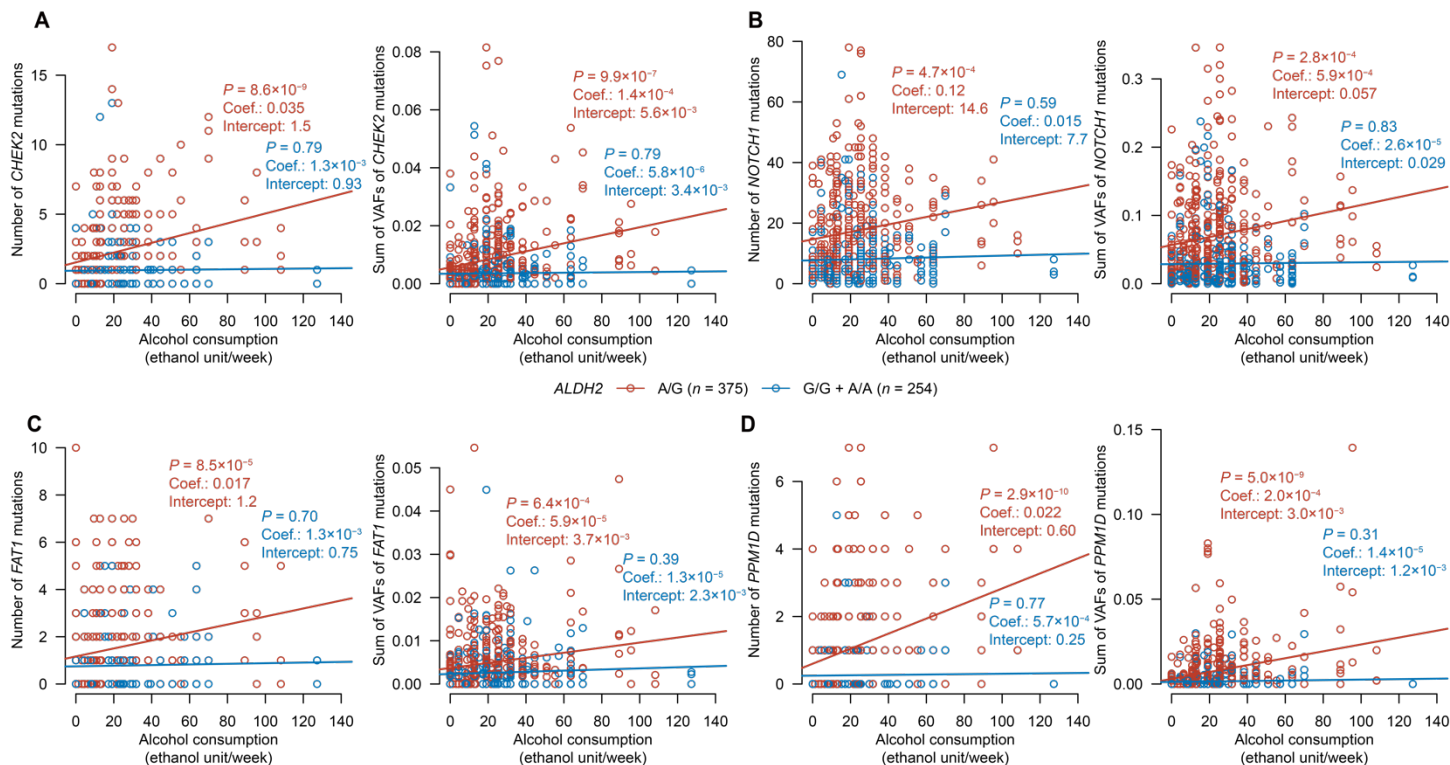


Fig. S7. Correlation of driver mutations to alcohol consumption. (A to D) Number (left) and sum of VAFs (right) of mutations in *CHEK2* (A), *NOTCH1* (B), *FAT1* (C), and *PPM1D* (D) plotted against alcohol consumption (ethanol unit/week) in samples from individuals with low- (blue) and high-risk (red) variants of *ALDH2*. The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model are indicated.

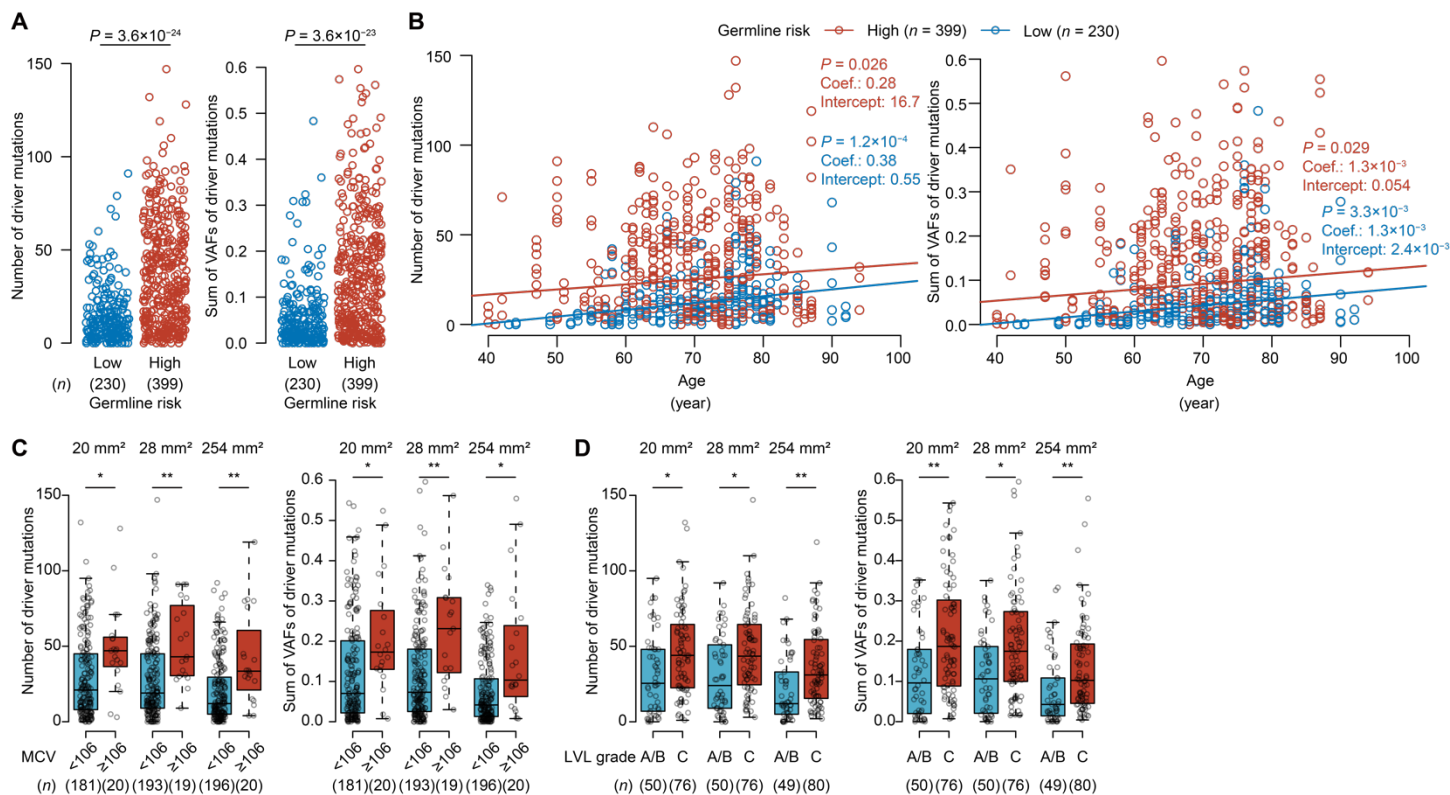


Fig. S8. Relationship of buccal mucosal remodeling with ESCC characteristics. (A) Number (left) and sum of VAFs (right) of driver mutations are plotted for solitary samples, stratified by germline risks. P-values for significance between groups are provided (Mann–Whitney *U* test). (B) Number (left) and sum of VAFs (right) of driver mutations plotted against age for samples from individuals with germline low- (blue) and high-risk variants (red). The p-value (*t*-test), coefficient, and intercept of the linear mixed-effects model for each risk group are indicated. (C) Number (left) and sum of VAFs (right) plotted according to MCV value (MCV <106 fL vs. MCV ≥106 fL). (D) Number (left) and sum of VAFs (right) plotted according to LVL grade (grades A/B vs. grade C). Box plots represent the median, first quartile, and third quartile, with whiskers extending to the furthest value within a 1.5× interquartile range. **P* < 0.01 and ***P* < 0.001 by Mann–Whitney *U* test.

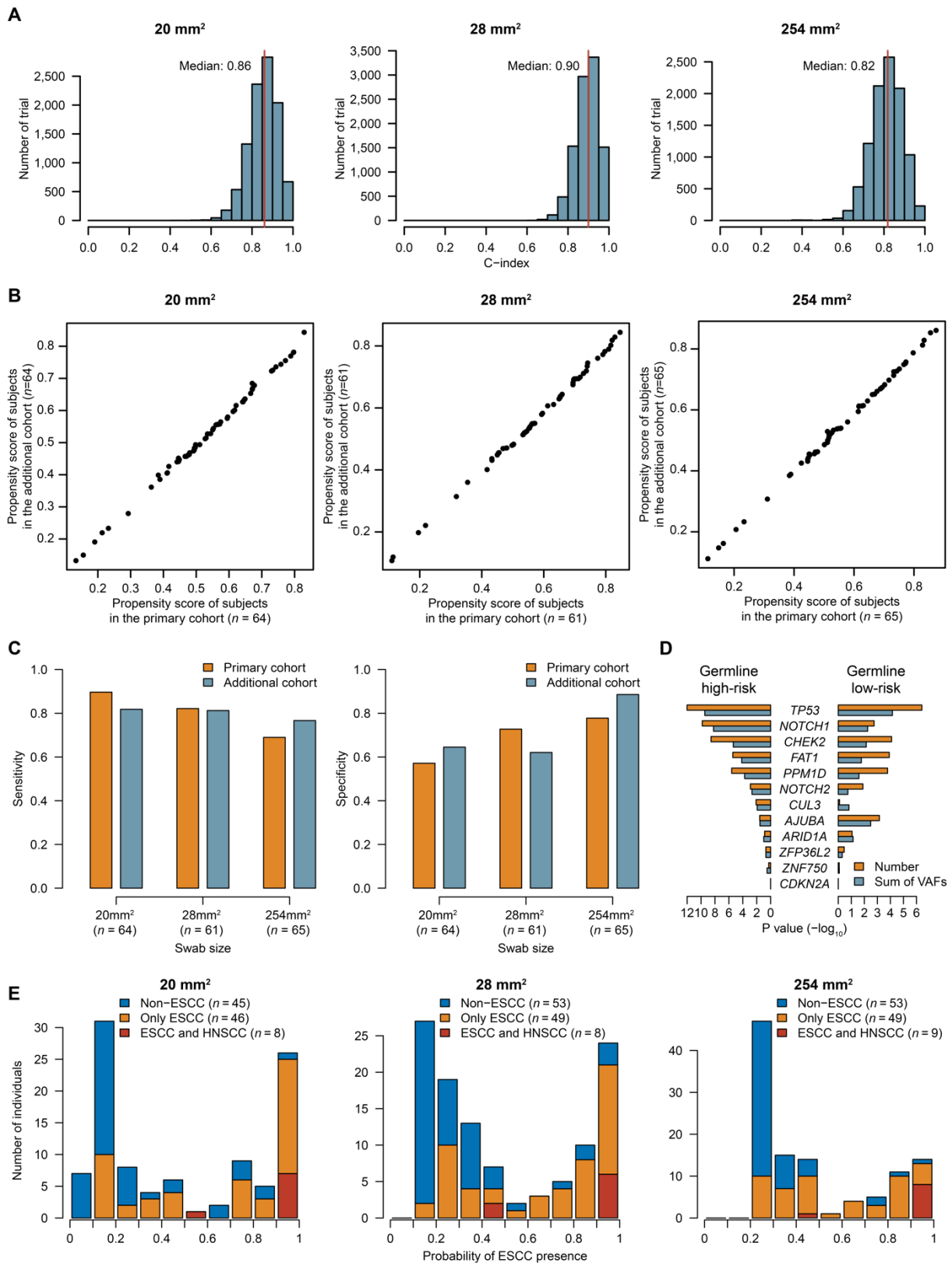


Fig. S9. Construction and validation of models to assess presence of ESCC. (A) Histograms of c-index calculated from ten thousand internal cross-validations for each swab type. The red line indicates the median value. (B) Propensity scores of selected subject pairs from the primary and additional cohorts. (C) Sensitivity (left) and specificity (right) of the primary cohort (orange) and the additional cohort (blue) are shown, using models constructed based on data from the primary cohort, with propensity score-matched samples assessed for both cohorts. (D) P values for the correlation of number (orange) or sum of VAFs (blue) of mutations in indicated

genes with the presence of ESCC, plotted for individuals with (left) and without (right) germline risk variants (Wald test). **(E)** Histograms of the probability of ESCC presence in the additional cohort, calculated by the models, are plotted for each swab type, with colors distinguishing individuals without ESCC, those with only ESCC, and those with both ESCC and HNSCC.

Table S1. Patient characteristics for the primary and additional cohorts.

Information on subject characteristics, including age, sex, sample number, germline variants status, alcohol flushing reaction, race, alcohol consumption, tobacco use intensity, mean corpuscular volume, cancer status, and grades of Lugol voiding lesions.

Table S2. Summary of patient characteristics.

Summary of characteristics of subjects with and without ESCC.

Table S3. Summary of consensus depth of each gene.

Median, minimum, and maximum consensus depth of bases within each analyzed gene.

Table S4. Results of digital droplet PCR experiments.

Results of a validation analysis using digital droplet PCR for randomly selected *TP53* mutations.

Table S5. Results of NanoSeq.

Number of mutations and total covered bases in NanoSeq experiments on 16 buccal swab samples.

Table S6. Patient characteristics for the analysis comparing tissue remodeling in the buccal mucosa and esophagus.

Information on subject characteristics, including age, sex, sample number, germline variants status, alcohol flushing reaction, race, alcohol consumption, tobacco use intensity, mean corpuscular volume, cancer status, and grades of Lugol voiding lesions.

Table S7. Results of dN/dS analysis.

Results of the analysis using dndscv on 26 genes.

Table S8. Driver genes of the esophagus and positively selected genes in the buccal mucosa.

Twelve out of 26 genes that exhibited positive selection in the buccal mucosa.

Table S9. Depth of error-corrected sequencing.

Information on raw depths, consensus depths, and final depths after downsampling, as well as the amount of input DNA.

Table S10. Depth of whole-exome sequencing.

Information on the average depth of whole-exome sequencing, as well as the amount of input DNA.

Table S11. Summary of linear mixed-effects models.

Metrics of linear mixed-effects models.

Table S12. Summary of logistic regression models.

Metrics of logistic regression models.