Supplementary Methods

Clinical Data Curation

TCGA Data

For TCGA study patients, the variable "DFS_STATUS" was used to identify patients with a clinical history of disease recurrence and/or progression. Values of disease-free status (DFS) in the dataset included unknown ("[Not Available]"), disease-free at the time of last follow-up ("DiseaseFree"), or recurred and/or progressed ("Recurred/Progressed"). Clinical TNM stages were used to define stage at diagnosis based on guidelines from a contemporary staging system for HPV-related OPSCC [1]. Primary tumors sites were categorized as either oral cavity or oropharyngeal as defined by the AJCC Cancer Staging Manual [2] using the "PRIMARY_SITE" variable. We utilized clinical TNM staging as pathological staging is not always available when primary treatment is non-surgical. HPV-unrelated tumors were assigned a clinical stage using the "CLINICAL_STAGE" variable. Smoking status was categorized as smoker or non-smoker where the non-smoker category was defined as ≤ 10 pack-years or "TOBACCO_SMOKING_HISTORY_INDICATOR" equals one (i.e., "Lifelong non-smoker" [3]). The variable "DFS_MONTHS" was used to define progression-free survival in downstream analyses.

Curation of TCGA, UW, and UPitt Molecular Data

Following mutation detection with MuTect2, we applied an additional set of filtering parameters to minimize the number of false positive mutations detected in our pipeline using parameters modified from prior work [4]. These parameters are illustrated in **Figure S2**. Briefly, we excluded mutants with less than 3X coverage or if the matched normal reference allele fraction was less than 97% suggestive of a potential germline variant. We also excluded somatic mutations variants in which the matched normal tissue had more than one variant allele

suggesting possible germline origin. We excluded variants with a tumor variant allele fraction less than or equal to 7% based on previous work illustrating moderate sensitivity for detecting somatic mutations with an 8% allele fraction using the MuTect algorithm [5]. Variants found on both strands were included in the final analyses. Variants whose population allele frequency in ExAC, ESP6500, or 1000Genomes was greater than 2% were excluded on the assumption that they were either present due to contamination, a variant present but missed in a normal sample, a low-level artifact, or that they would be unlikely to be a cancer driver mutation given the relatively frequent prevalence in the population [4,6,7]. Further exclusions were based on modified analytic methods used in the TCGA pancancer analyses [8] involving the exclusions illustrated in **Figure S3**. The following details describe the number of variants filtered by each parameter for each study resulting in the final dataset used for analyses in relation to recurrence status.

TCGA Data

A total of 513 MAF files were imported from the TCGA head and neck squamous cell data archives at FireBrowse (http://firebrowse.org/?cohort=HNSC&download_dialog=true) containing 120,398 observations from 510 unique tumor samples. After limiting the dataset to observations with a tumor alternate allele count of greater than or equal to three, there were 119,417 observations remaining. A variable representing the tumor variant allele frequency (VAF) was generated by dividing the tumor alternate allele count by the sum of the tumor alternate and reference allele counts. We filtered the data using a tumor VAF threshold of greater than or equal to seven percent resulting in 106,877 remaining observations. Possible germline variants were excluded by removing observations with an allele frequency of > 0.02 in the ExAC, ESP6500, or 1000Genomes databases with 106,803 observations remaining. Variants with a Hugo symbol beginning with “LOC” or “ENSG” were excluded eliminating 222 variants. Variants known to be problematic on GRCh37 or with a propensity for mutation based on size including *PE4DIP*, *CDC27*, *MUC4*, *DUX4*, *TTN*, *HYDIN*, *PRIM2*, *MUC16*, *OBSCN*,

AHNAK2, *SYNE1*, *FLG*, *MUC5B*, *DNAH17*, *PLEC*, *DST*, *SYNE2*, *NEB*, *HSPG2*, *LAMA5*, *AHNAK*, *HMCN1*, *USH2A*, *DNAH11*, *MACF1*, and *MUC14* were excluded eliminating an additional 2,112 variants from the analysis, leaving a total of 104,469 variants. Molecular filtering parameters led to the exclusion of four samples including three HPV-related OPSCC tumors from non-recurrent cases and one from a recurrent case. Among the 36 TCGA HPV-related primary OPSCC study participants, there were a total of 3,665 SNVs and indels between 29 tumors that did not recur and seven that did recur.

Universities of Washington and Pittsburgh Data

UW and UPitt whole exome sequencing data were obtained as described in the **Methods**. Raw data were processed in the same manner as the TCGA data with regards to sequencing alignment, recalibration, and variant calling using GATK Best Practices and MuTect2.0. Once the processed data were obtained, we performed post-processing as described for the TCGA data in the previous section. Among the 21 HPV-related primary OPSCC study participants from UW and UPitt, there were a total of 26,831 SNVs between 12 tumors that recurred and 9 tumors that did not recur.

Defining Driver Genes

We defined a list of 467 genes for which mutations critical to tumorigenesis and/or that are potentially targetable have been described. This list was comprised of genes based on prior work of Vogelstein et al. [11], Hedberg et al. [12], and Pritchard et al. [13] In order to capture additional genomic aberrations central to tumorigenesis, we also included genes involved in cell cycle control, cell death, survival, DNA damage response, PI(3)K signaling, RTK signaling, Ras-Raf-MEK-Erk/JNK signaling, and differentiation [based on gene lists in the cBioPortal database [10,11]].

Copy Number Analysis

We utilized ADTEx (Aberration Detection in Tumour Exome) which normalizes coverage between normal and tumor samples to infer somatic copy number changes among our study population [15]. ADTEx implements a circular binary segmentation algorithm from the Bioconductor package, DNAcopy [16]. Segmented copy number ratios are log2 transformed and prepared for input into GISTIC. GISTIC2.0.22 was implemented on GenePattern [17] to identify significantly amplified or deleted regions across the genome among the primary tumors of HPV-related OPSCC tumors that did or did not recur [18-20]. Input parameters used were as described above for TCGA head and neck squamous cell data pipeline including amplifications and deletions thresholds = 0.1, joint segment size = 4, q-value threshold = 0.25, confidence level = 0.99, cap value = 1.5, focal length cutoff = 0.70, max sample segments = 2000, arm peel = 1, sample center = median, and gene collapse method = extreme.

References

1. Dahlstrom KR, Garden AS, William WN, Lim MY, Sturgis EM. Proposed staging system for patients with HPV-related oropharyngeal cancer based on nasopharyngeal cancer N categories. *J Clin Oncol.* 2016;34(16):1848-1854.
2. Edge S, Byrd D, Compton C, Fritz A, Greene F, Trotti A. *AJCC Cancer Staging Manual, 7th Edition.* France: Springer; 2010.
3. Lee H, Palm J, Grimes SM, Ji HP. The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations. *Genome Med.* 2015;7:112.
4. Ho AS, Kannan K, Roy DM, et al. The mutational landscape of adenoid cystic carcinoma. *Nat Genet.* 2013;45(7):791-798.
5. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics.* 2014;15:244.
6. Hakenberg J, Cheng WY, Thomas P, Wang YC, Uzilov AV, Chen R. Integrating 400 million variants from 80,000 human samples with extensive annotations: towards a knowledge base to analyze disease cohorts. *BMC Bioinformatics.* 2016;17:24.
7. Pandya C, Uzilov AV, Bellizzi J, et al. Genomic profiling reveals mutational landscape in parathyroid carcinomas. *JCI Insight.* 2017;2(6):e92061.
8. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature.* 2013;502(7471):333-339.
9. Morris LG, Chandramohan R, West L, et al. The molecular landscape of recurrent and metastatic head and neck cancers: insights from a precision oncology sequencing platform. *JAMA Oncol.* 2016.
10. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 2013;6(269):pl1.
11. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2012;2(5):401-404.
12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science.* 2013;339(6127):1546-1558.
13. Hedberg ML, Goh G, Chiosea SI, et al. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. *J Clin Invest.* 2016;126(1):169-180.
14. Lui VW, Hedberg ML, Li H, et al. Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discov.* 2013;3(7):761-769.
15. Amarasinghe KC, Li J, Hunter SM, et al. Inferring copy number and genotype in tumour exome data. *BMC Genomics.* 2014;15:732.
16. Seshan V, Olshen A. DNAcopy: DNA copy number data analysis. *R package version 1501.* 2017.
17. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Genet.* 2006;38(5):500-501.
18. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 2011;12(4):R41.
19. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature.* 2010;463(7283):899-905.
20. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A.* 2007;104(50):20007-20012.