S1. Biospecimen collection and clinical data

S1.1. Biospecimen collection and clinical data

Overview

Smoking is a primary risk factor for HNSCC, and the rise in smoking rates in developing countries contributes to the increased prevalence of the disease worldwide. In addition, HPV infection is now recognized as an important factor for oropharyngeal tumors among non-smokers. Despite surgery, radiation, and chemotherapy, approximately half of all tumors will recur either regionally or with distant spread, usually within two years of initial diagnosis. The prognosis for recurrent cancer is dismal.

In the text we present each data type separately to define somatic alterations and molecular subtypes. Subtypes are categorized with the intention to emphasize somatic variants enriched in subpopulations rather than as definitive diagnostic categories. In selected cases, combinations of data types are considered (i.e. mutation and copy number alteration), where data integration highlights the importance of a target, coordinated somatic alterations that highlight pathway dysregulation, and characterization and/or validation of molecular subtypes. Using this approach, we identify novel and known alterations that elucidate key pathways, networks, and subtypes of potential biologic, prognostic, and therapeutic interest in HNSCC.

Sample Acquisition

Tumor samples were accrued as part of The Cancer Genome Atlas (TCGA) network and were obtained from patients with appropriate consent from the relevant Institutional Review Board. Briefly, tumors were resected, flash-frozen, and shipped to a centralized processing center (Biospecimen Core Resource, BCR) for additional pathologic review and extraction of nucleic acids. Aliquots of DNA and RNA were shipped to individual sites for all subsequent testing. Normal DNA samples were provided as processed DNA.

Biospecimens were collected from newly diagnosed patients with Head and Neck Squamous Cell Carcinoma (HNSCC) at the time of surgical resection. Unless otherwise noted, the patients had received no prior treatment for their disease including chemotherapy or radiotherapy. Similar to clinical populations almost all patients were treated with curative intent [1]. All cases were collected without regard to surgical stage or histologic grade. Cases were staged according to the American Joint Committee on Cancer (AJCC), Seventh Edition. Each frozen tumor specimen had a companion normal sample which consisted of blood components, adjacent normal tissue more than 2 cm from the tumor, or previously extracted germline DNA from blood or normal tissue. Each frozen tumor specimen submitted to the BCR weighed at least 30 mg and was typically under 200 mg. Specimens were shipped overnight from one of 14 tissue source sites using a cryoport that maintained an average temperature of less than -180°C. The tissue source sites contributing biospecimens included: Asterand, Inc.; GPCC, Greater Poland; International Genomics Consortium; Indiana ABS; Johns Hopkins University; MD Anderson; Medical College of Georgia; Roswell Park Cancer Institute; University Health Network; University of Miami; University of Michigan; University of North Carolina; University of Pittsburgh; and Vanderbilt University.

Clinical Data

Complete clinical data elements were collected for all specimens to include: sample code, primary site (head and neck sub-site for all specimens), gender, age at diagnosis, race, ethnicity, year of tumor collection, laterality (left, right, midline), tumor grade, smoking and alcohol history, and HPV status. Risk stratification for HNSCC is currently limited to anatomic site, stage, and histologic characteristics of the tumor [2]. Except for HPV status, numerous molecular and clinical risk factors have been investigated without validation for potential clinical application [3]. For this reason, few patients had documentation for any molecular biomarkers. Patients with a history of a prior malignancy and/or prior therapy had relevant data annotated.

In addition, Tumor, Node, and Metastasis (TNM) staging components were requested for both clinical and pathologic staging for all cases in the study. A compiled tumor stage using the standard AJCC staging criteria for TNM was reported. Clinical and vital (living/deceased) status as well as tumor status (with tumor or tumor free following tumor resection) of patients at the point of enrollment and, as available, at last follow-up was also recorded.

Follow-up data were requested for subjects from the time of sample collection. The patient's tumor status (tumor free/with tumor) was again recorded, along with vital status (living/deceased) from the most recent follow-up data form completion. Time to recurrence was calculated as the number of days to a new tumor event. Study follow up mirrored the typical clinical pattern in which recurrences are within two years of diagnosis [1], and the prognosis for these patients have poor outcomes overall [4]. The number of days to last contact at the point of enrollment or most recent follow up was also recorded. Finally, the days from diagnosis (sample collection) to death was recorded at both enrollment and in the most recent follow up forms. These data provide the information to explore survival-based outcomes and median follow-up for patients included in this study. Cause of death, from cancer or other causes, was not recorded.

Smoking Status

For the purposes of all analyses, smoking status was coded as a binary term. Patients who were reported as lifelong nonsmokers or who had reported pack years of less than or equal to 10 years were considered light smokers or non-smokers. All other subjects were considered smokers. For the sake of analysis, the few patients missing smoking status at the time of the data freeze were categorized as smokers.

Verification of Histologic Diagnosis

Tumors selected met the histologic criteria for squamous cell carcinoma. Each tumor and adjacent normal tissue specimen were embedded in optimal cutting temperature (OCT) medium and histologic sections were obtained and stained with hematoxylin and eosin (H&E) stain for review. Each H&E image from the frozen section was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with HNSCC and that the adjacent normal specimen contained no tumor cells. The tumor sections were required to contain an average of 60% tumor cell nuclei with less than 20% necrosis for inclusion in the study per TCGA protocol requirements. Eight of the tumors that previously were disqualified due to insufficient tumor cell nuclei or greater than 20% necrosis were recovered with Laser Capture Enrichment (LCE). LCE is a proprietary, high throughput technology that allowed the BCR to accurately macro/micro-dissect tumors in order to isolate tumor within frozen sections with no chemical staining or fixing of tissue (avoiding RNA or DNA potential effects).

Biospecimen Processing

RNA and DNA were extracted from tumor specimens using the Allprep DNA / RNA Kit (Qiagen) in accordance with the mirVana miRNA Isolation Kit (Ambion). This protocol generated RNA preparations that included RNA <200 nt [designated 'mirVana (Allprep DNA) RNA'] suitable for miRNA analysis. DNA was extracted from normal controls using either the QiaAmp blood midi kit (Qiagen) or the combination of the Allprep DNA / RNA kit (Qiagen) and the mirVana miRNA Isolation Kit (Ambion) as stated above with the tumor specimens. DNA specimens were quantified by Picogreen fluorescence assay and resolved by agarose gel electrophoresis to determine the range for fragment sizes. The Sequenom iPLEX sample identification panel was utilized to verify that tumor DNA and germline DNA were derived from the same patient. One µg each of tumor and normal DNA was sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only those specimens yielding a minimum of 6.9 µg of tumor DNA, 4.9 µg of germline blood DNA, or 6.9 µg solid tissue normal DNA, and 5.15 µg of tumor RNA, were included in this study. RNA was quantified by measuring Abs260 with UV spectrophotometer, which was then analyzed via the RNA6000 assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN > 7.0 were included in this study.

Data Freeze

For the purposes of performing an integrated analysis across multiple platforms, the HNSCC analysis working group established a "data freeze" set of samples. Tumors were included in the data freeze if they had a complete set of molecular assays as of 9/19/2012, the date of a face-to-face workshop held at the Lineberger Comprehensive Cancer Center at the University of North Carolina at Chapel Hill. A total of 279 tumors were included with clinical data, whole exome sequencing (WES), RNA sequencing, miRNA sequencing, methylation arrays, and DNA SNP chips. Additional assays were included for a subset of patients. In the majority of cases patients had matched normal peripheral blood mononuclear cells as a source of DNA for matched assessment of somatic mutations in tumors and for the purposes of SNP chip normal control array. In selected cases where DNA from blood was not available or the resulting assay did not pass quality standards, tumor adjacent normal was evaluated. A subset of 37 samples had additional tumor adjacent normal tissue (Section S5.3) which was then assayed on the miRNA and RNA platforms. A subset of 50 samples were used as normal controls for the DNA methylation platform, 20 of which overlap with the 37 normals available for miRNA and mRNA. Ninety-eight of the 279 data freeze samples had "low pass" whole genome sequencing (WGS). Whole genome sequencing was performed on 29 of 279 samples. Reverse phase protein arrays (RPPA) were performed on 200 of 279 samples. An accounting of all samples and normal controls used is available in Data File S1.1. When viewing figures containing data from the two whole genome platforms or RPPA (e.g. Figure 3, where phosphorylation data are missing for some cases and Figure S3.10 where not all HPV positive cases had whole genome sequencing) it should be noted that some cases with absent findings may represent missing data for the specific platform.

In general, all genomic data from the data freeze were processed through standard TCGA analytic pipelines and are accessible through TCGA data portals including https://tcga-data.nci.nih.gov/tcga/ and https://tcga-data.nci.nih.gov/docs/publications/hnsc_2014/. In cases where versioned subsets of TCGA data are specific to the data freeze and the current analysis, locked versions of the files such as the clinical data file (Data File S1.1) are also made available through the Data Coordinating Center. Also available at the link are results from a limited number of analyses which were not a routine component of TCGA data archives at the time of publication. Examples of such analyses include some structural alteration DNA and RNA, particularly those that reference non-human (virus) sequencing data and the HPV status of each case (HPV(+) or HPV(-), Data File S1.2). Additionally, manual review of clinical data revealed some inconsistencies across tissue source sites which were resolved by clinical expert review. The clinical files used for the current analysis are available at the TCGA data portals above.

S1.2. HPV detection methods

Clinical HPV Classification

Using a convention established by a landmark clinical report on HPV in clinical samples, we defined a patient as "clinically positive" for HPV when the following conditions were met: 1) the tumor was located in an oropharyngeal sub-site (base of tongue, tonsil, soft palate, or oropharynx not otherwise specified), and 2) the patient had a positive assay for HPV performed as part of routine clinical care and reported in the electronic case report forms (Figure S1.1) [5]. The assay was recorded as either p16 immunohistochemistry staining or *in situ* hybridization and the result recorded as either positive or negative. No further details were collected as to the reagents used at the individual tissue source sites. In cases where oropharyngeal tumors did not have HPV testing performed or reported through electronic case report forms, we requested that separate paraffin sections be supplied for centralized HPV testing as described below. Only oropharyngeal tumors with no HPV status were requested, although a limited number of non-oropharyngeal samples were received and tested. Oropharyngeal tumors testing positive after central review were also classified as "clinically positive." Oropharyngeal tumors that were not received for review and that were missing data from their case report forms were classified as "HPV missing." All tumors from a non-oropharyngeal site were classified as "clinically HPV negative" by convention although we considered other definitions of HPV positive, including "HPV molecular positive" as discussed below.

p16 Immunohistochemistry

Unstained sections or tumor blocks from oropharyngeal tumors with no clinical assay for HPV were requested. 4-µm thick sections were prepared from a representative paraffin block of the respective carcinoma and were incubated at 60 degrees for twenty minutes. Following heat induced epitope retrieval with Citrate buffer for 15 minutes at 100 degrees; slides were incubated with antibody p16 (clone E6H4) (Ventana Medical Services); 1:3. The Refined Polymer Detection kit was used for detection of bound antibody, with 3,3-diaminodenzidine serving as the chromagen (Leica Microsystems). Slides were counterstained with Mayer's hematoxylin.

Scoring: p16 was scored in binary fashion as positive when strong homogeneous cytoplasmic and/or nuclear staining was found and negative if no staining or focal heterogeneous nuclear or cytoplasmic staining was observed.

In situ hybridization

Unstained 4 µm thick sections were processed from formalin fixed paraffin embedded tumor blocks for high risk HPV subtypes (#16, #18, #33, #35, #39, #45, #51, #52, #56, #58 and #66) which detects the Inform HPV III family 16 probe (Ventana Medical Systems) in conjunction with ISH IV (VIEW) Blue Plus detection kit on the Bench Mark series automated slide stainer (Ventana Medical Systems). Sections from HPV positive oropharyngeal cancer were used as a positive control.

Scoring: The observation of single or multiple punctate hybridization signals localized to tumor cell nuclei was scored positive. The lack of any hybridization dots was considered negative.

HPV detection by multiple nucleic acids techniques

Assessment of HPV using all nucleic acids-based sequencing platforms included RNA-Seq (n=279 samples, S3), WES (n=279 samples, S4), "high coverage" WGS (n=29, S4), and "low pass" WGS (n=98, S4). The methods for each platform are described below.

HPV detection by RNA-Seq

Expression status and viral gene expression were determined by sequence alignment by Mapsplice[6]. All reads that remained unaligned after an initial alignment to the human genome were identified and subsequently aligned to a database of microbial genomic sequences and to a publicly available database of viral sequences. Reads mapping to the viral genome were then quantified using the Bioconductor package of the R statistical computing language GenomicFeatures[7,8]. Mated reads were counted once. In cases where the database of viral sequences contained multiple reference sequences attributed to a single named strain, a final strain attribution was based on the strain to which the greatest number of reads aligned. Using this method, a low rate of false positive alignments to viral sequences was observed which was largely attributed to homology to the human genome in falsely mapped sequences.

Only strains of HPV were detected at high aligned read counts with the exception of a single high-coverage case of herpes simplex virus 1 (HSV1) in sample TCGA-BA-5558. Interrogation of clinical material from TCGA-BA-5558 revealed abundant tumor in paraffin sections with negative staining for HSV1 by clinical immunohistochemical staining with positive controls (data not shown). A clinical qualitative PCR assay for HSV1 in TCGA-BA-5558 was weakly positive with the conclusion that a rare cell was likely positive for the virus although the tumor overall contained little HSV1.

Among samples with sequence reads aligning to HPV a striking distribution was observed (Figure S1.1). Thirty-six of 279 cases demonstrated a large number of reads ("high coverage", greater than 1000 reads), with clear evidence of expression of some or all viral genes including E6 and E7. High-coverage cases overwhelmingly corresponded to the clinical and molecular picture of HPV infection including oropharyngeal site, non-smoking history, and lack of mutation in the gene *TP53* (Figure S1.1). For these reasons, samples were classified as HPV positive using an empiric definition of detection of > 1000 mapped RNA-Seq reads, primarily aligning to viral genes E6 and E7 (Figure S1.1). In some cases, samples documented positivity for multiple strains, possibly due to viral homology, in which case the sample was assigned to the HPV subtype with the greatest read count. An additional set of approximately 60 samples documented aligned read counts between 1-33 reads, possibly representing latent infection as opposed to false positive. HPV alignments of even a single read were rare in matched normal controls and in samples obtained from non-HPV associated cancers evaluated by TCGA. The clinical and biological relevance of low coverage HPV-containing samples is unclear.

HPV detection by "high coverage" whole genome sequencing and whole exome sequencing data

The PathSeq pipeline was used to perform computational subtraction of human reads, followed by alignment of residual reads to a combined database of human reference genomes and microbial reference genomes (which includes but is not limited to HPV genomes), resulting in the identification of reads mapping with high confidence to HPV genomes in WGS and WES sample data. Subjects were classified as HPV positive by WGS if the HPV read counts were at least 100; subjects were HPV positive by WES if the HPV read counts were positive at any level (even one read); otherwise subjects were classified as HPV negative as was previously published [9]. In addition to identification of HPV strain type, whole genome sequencing data were also analyzed to identify HPV integration sites. In contrast to prior studies, mutation rates did not differ by HPV status, although transversions at CpG sites were more frequent in HPV(-) tumors while a predominance of TpC mutations were noted in HPV(+) cases (Figure S1.1) [10].

In brief using the PathSeq pipeline, human reads were subtracted by first mapping reads to a database of human genomes using BWA, MegaBLAST and BLASTn. Only sequences with perfect or near perfect matches to the human genome were removed in the subtraction process. To identify HPV reads, the resultant non-human reads were aligned to a database of microbial genomes that includes multiple HPV reference genomes with MegaBLAST and BLASTn. HPV reference genomes were obtained from the Human Papillomavirus Episteme [11]. Chimeric human and HPV read pairs were identified by extracting the pair mates of HPV reads and aligning the paired end (PE) reads to a combined human and HPV reference genome, using BWA. The chimeric read pairs, in which one read maps to the human genome and the mate maps to the HPV genome, represent integration sites. Then, high confidence (at least 3 spanning reads that cover the integration event) putative HPV integration sites were identified in the samples.

HPV detection by "low pass" whole genome sequencing

To detect viruses and to examine the physical status of the viral genome in HNSCC WGS DNA data, an in-house-developed pipeline was used. As the first step, the pipeline performed computational subtraction of sequences mapped previously to the human genome. Then it used BWA aligner to map the remaining set of non-human sequences to the set of viral reference genomes obtained from the NCBI RefSeq database [12].

Reads that aligned to the genomes of multiple species were filtered out. Percentage of covered viral genome, count of virus sequencing reads normalized by the length of the viral genome, and total number of non-human reads in the sample was calculated. To consider a given sample positive for the presence of the virus we chose an empirical threshold of 100 viral read counts.

To assess virus integration into the host genome the pipeline used the advantage of paired-end (PE) sequencing technology and searched for the clusters of discordant read pairs where one mate is aligned to the human genome and the second mate mapped to the viral sequence. As an input original set of all PE reads, those mapped and unmapped to the human genome were used and two subsets of reads were generated: ends represented by human sequences and their unmapped mates. Then such unmapped reads were aligned against the specific viral genome identified in the previous step. Clusters of discordant read pairs were calculated. To determine the presence of a cluster we used an empirical cutoff of 3 discordant read pairs within the same integration region. To assess the precise site of integration at nucleotide resolution the pipeline searched for the chimeric viral-human reads. Soft-clipped reads where only part of a read was mapped to the human genome were filtered from the original PE dataset and were aligned by BLAT to the virus genome.

HPV detection by MassArray

Recognizing that by convention most samples in the data freeze did not have clinical assays performed for HPV due to low sensitivity and specificity of p16 and *in situ* hybridization outside orpharyngeal sites, and that many of the genomic platforms above would have a limited track record in clinical cases, the analysis working group recommended at least one additional assay be considered for all samples. With this goal HPV status in TCGA qualified HNSCCs was evaluated by the Biospecimen Core Resource at Nationwide Children's Hospital (TCGA Center 23). A PCR based MassArray assay developed at the University of Michigan was performed to test a panel of 16 HPV types (#16, #18, #31, #33, #35, #39, #45, #51, #52, #56, #58, #59, #66, #68, #73, and #90). Multiplex PCR and Sequenom-based Mass Spectrometry was used to detect the presence of specific HPV types based on the DNA sequence heterogeneity found within individual viral E6 regions. Tumor DNA was classified as HPV(+) or (-) based on the reproducible detection or failure to detect one of the HPV types included in the panel. If positive for a single HPV type or more, the specific HPV type identified is specified in the "MassArray_HPV_Call" column of Data File S1.2. Tumor DNA that did not reproducibly detect the presence of HPV was classified as "Indeterminate."

HPV prediction by molecular signatures: miRNA signature for HPV status

In addition to direct assessments of HPV status, multiple groups have previously reported "HPV signatures" including an miRNA signature strongly associated with clinical HPV positivity. In brief, miRNAs that are differentially expressed ($q<0.01$) between HPV(+) and HPV(-) oropharyngeal carcinomas (OPCs) were identified using miRNA-Seq count data, TMM normalization, tagwise dispersion, and a classic exact test in edgeR v3.0.8 [13]. A panel of 5 differentially expressed miRNAs (miR-20b, miR-9, miR-146a, miR-193b, and miR-363) that had been previously described to be associated with HPV in OPC[14-17] was selected for further analysis. Using the 33 OPC samples in the data freeze sample set, (22 HPV(+), 11 HPV(-)) a signature associated with HPV positivity was generated by fitting the RPM expression profiles of the 5 miRNAs to a logistic regression model (glm() function, R v2.15.2). A risk score was assigned to each patient using the formula generated from the model (below). The median risk score for OPC samples (5.56) was chosen as a binary cut-point and was applied across all HNSCC samples (n = 279).

Using the coefficients from the model, the following formula was generated: risk score = RPM(miR-20b)·0.72 + RPM(miR-146a) ·$7.61\text{x}10^{-2}$ - RPM(miR-9)·$8.75\text{x}10^{-4}$ - RPM(miR-193b)·0.55 - RPM(miR-363)·0.31. Figure S1.2 shows the risk score's predictive ability in OPC samples compared to that of each of the miRNAs individually. When all HNSCC samples were dichotomized based on their risk scores being equal to or greater than vs. less than the median for OPC samples (miRNA track of Figure S1.1), the resulting binary miRNA score had a sensitivity and specificity of 0.56 and 0.96, respectively. When compared to two other molecular indicators of HPV status such as *CDKN2A*-loss (sensitivity = 1, specificity = 0.73) and *TP53*-mutation (sensitivity = 0.97, specificity = 0.76), the miRNA signature is less sensitive but more specific.

HPV prediction by molecular signatures: methylation signature for HPV status

Extending the observation that genomic signatures may play a role in identification of HPV positive samples, we used empirical Bayes modified t-tests as implemented in the limma package [18] in R to identify CpG probes that were correlated with HPV status, while adjusting for the confounding effects of tissue type. The association between HPV status and DNA methylation was strong enough to support very strict criteria for probe selection, signature genes were selected to have false discovery rate (FDR) less than 1.0e-10 (q < 1.0e-10). CpG sites at which HPV(+) samples show significantly increased methylation were averaged after centering each probe to the mean of the HPV(-) samples to create an HPV(+) DNA hypermethylation score. An HPV(+) hypomethylation score was similarly computed from sites correlated in the other direction. A sample was called HPV(+) if both the hyper- and hypomethylation HPV scores exceeded thresholds defined on the training set. The entire procedure, from probe selection to defining of the thresholds for the calls, was embedded in a leave-one-out cross-validation so that we could make unbiased calls on the training samples as well as the suspicious samples.

CpG probes that were significantly correlated with HPV status were combined into two signatures, depending on whether they were hyper- or hypomethylated in HPV(+) samples. Results are shown in the left-most panel of Figure S1.3. To test the signature we calculated scores for TCGA cervical samples (middle panel) and for an independent set of HNSSC tumors (left panel) [19]. Our signature differed slightly from the one derived by Lechner et al. [19] using the samples shown in the panel on the right, in that we found significant HPV-associated changes in both directions, while they found HPV(+) samples to largely gain methylation. The same study posited a role for the polycomb repressive complex in the HPV hypermethylation signature. We found 51/108 or 47% of the genes in the hypermethylated signature to be polycomb, a typical level of enrichment for hypermethylated genes.

HPV detection summary

A summary of the HPV detection methods is presented in Figure S1.1 and Data File S1.2. As noted above, samples were classified as HPV positive using an empiric definition of detection of > 1000 mapped RNA-Seq reads, primarily aligning to viral genes E6 and E7 (Figure S1.1). In cases where subjects had positive read counts for more than one HPV strain, the strain was defined by the highest read count. In general there was good agreement between all measures of HPV assessment, although some patterns are worth noting. Although 64% of the 33 oropharyngeal tumors were HPV(+), 14 of the 36 cases in which the molecular data strongly support an HPV(+) diagnosis were from anatomic sites for which HPV testing is not indicated. This suggests that while the rate of HPV(+) cases from non-oropharyngeal sites may be low at 6% (14/246), a significant burden of HPV(+) disease may reside outside the oropharynx [20]. Additionally, the data support our belief that sensitive methods of HPV detection such as low abundance RNA coverage or sensitive DNA assays such as MassArray will detect higher rates of HPV including episomal virus that is not considered causative of the tumor. Lower smoking rates for HPV(+) patients were observed on average, however most HPV(+) patients were smokers. Evaluation of the types of mutations seen vary as a function of HPV status as measured by the nucleotide score. Using a measure of enrichment for transversions defined for each subject as the log2 of the ratio of the number of (TpC to G/T) events to the number of the (CpG to T) events, higher scores were evident in the HPV positive samples. As expected, p16 inactivation (mutation, methylation, or homozygous deletion of *CDKN2A*) as well as *TP53* mutations were essentially limited to HPV(-) tumors.

Mutation signatures by HPV status

In their analysis of sequencing data from over 7000 primary tumors, Alexandrov et al.[21] identified 21 mutational signatures and determined which of these signatures were present in 30 distinct tumor types. The authors found that HNSCC tumors exhibited four mutational signatures, including the APOBEC and smoking signatures. These were characterized by C > T and C > G mutations in TpCpN trinucleotides and C > A mutations, respectively. Our analysis showed a remarkably strong association between these mutational signatures and HPV status (Figure S1.1, Data File S1.3). The HPV(+) samples displayed a significant enrichment for the APOBEC signatures (Fisher's exact test $p < 2.2e-16$), while the smoking signature was pronounced in the HPV(-) samples (Fisher's exact test $p < 2.2e-16$).

S1.3. Survival analysis

The survival R package [22] was used to analyze overall survival times, produce Kaplan-Meier plots, and compute log-rank test p-values, as shown in Figures S1.4 and S1.5. Overall survival times were censored at 60 months given that most cancer-related events occur before that time. Tumor stage was defined to be the surgical stage. HPV status was determined by RNA sequencing, as defined above. Subjects were categorized as having 11q13 or *EGFR* amplifications if their discrete GISTIC copy number value for *CCND1* or *EGFR*, respectively, was equal to two, as described in S2.1, otherwise they were considered to have no amplification. *TP53* mutation status was assessed by exome sequencing, as described in Section S4. Tumor site was classified as oral cavity if the tumor sample came from any of the following anatomic subdivisions: buccal mucosa, floor of mouth, hard palate, lip, oral cavity, oral tongue, and alveolar ridge; tumor site was classified as oropharynx if the tumor sample came from either tonsil or oropharynx. Hierarchical clustering of DNA copy number data was used to identify DNA copy classes, as described in S2.1. Copy number class 3, denoted somatic copy alteration (SCNA) quiet, was enriched for HPV(+) samples (25/72). The HPV(-) subjects in copy number class 3 demonstrated remarkably better survival outcomes when compared to the remaining HPV(-) subjects, as shown in Figure S1.5. These observations provide confirmation of the preliminary findings of Smeets et al[23].

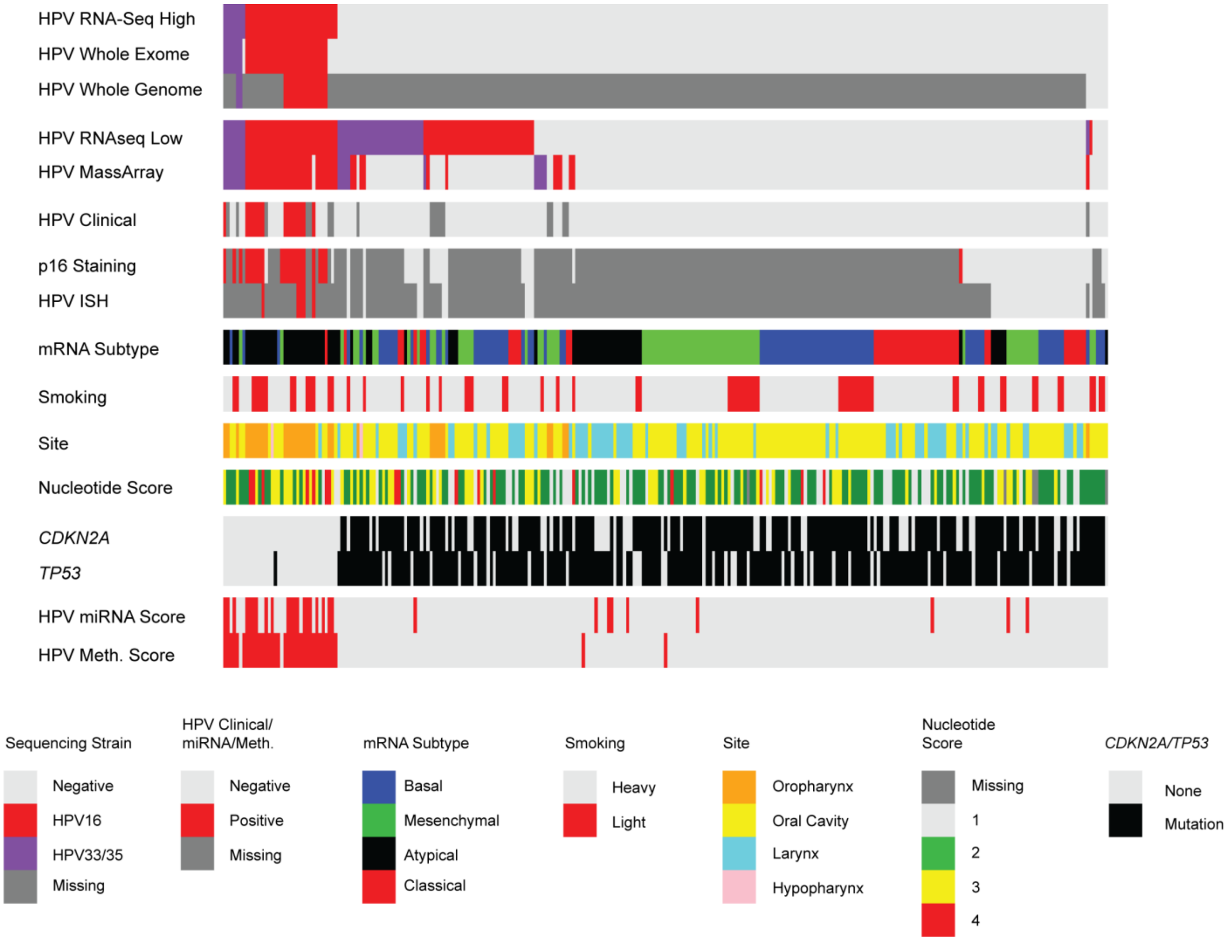


Figure S1.1, part 1. HPV status as a function of clinical and molecular characteristics. Samples are arranged in columns and grouped according to HPV status, as detected by RNA, whole exome, and whole genome sequencing. Samples were classified as HPV positive , as shown in the “HPV RNA-Seq High” annotation track, using an empiric definition of at least 1000 mapped RNA-Seq reads, primarily to the viral genes E6 and E7. Additional annotation tracks provide summaries of alternate HPV detection methods, and these are described in S1.2. ISH, *in situ* hybridization; Meth, methylation

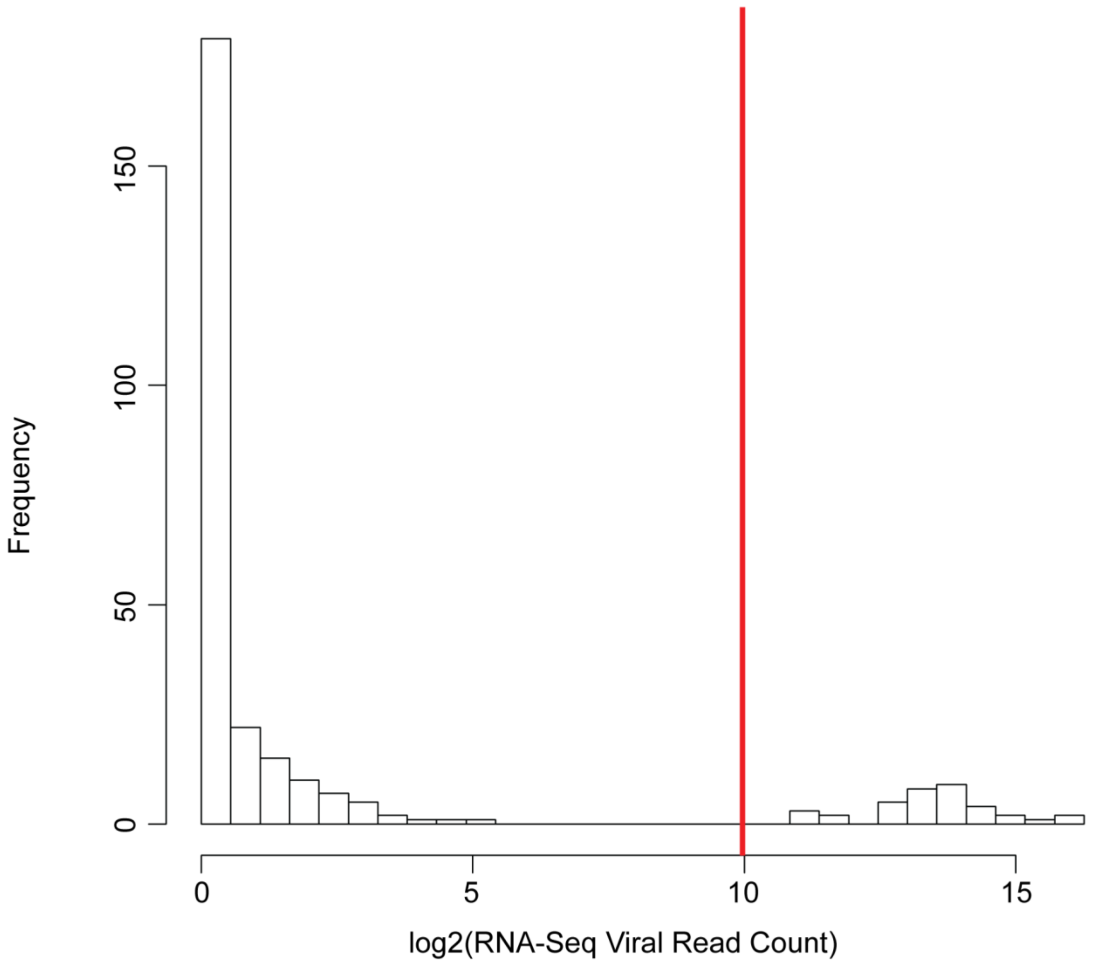


Figure S1.1, part 2. Distribution of RNA-sSq viral read counts. A histogram of log2(RNA-Seq viral read counts) for the 279 HNSCC samples exhibits a bimodal distribution. HPV status was defined based on an empiric definition of at least 1000 mapped reads. This threshold is illustrated by the vertical red line.

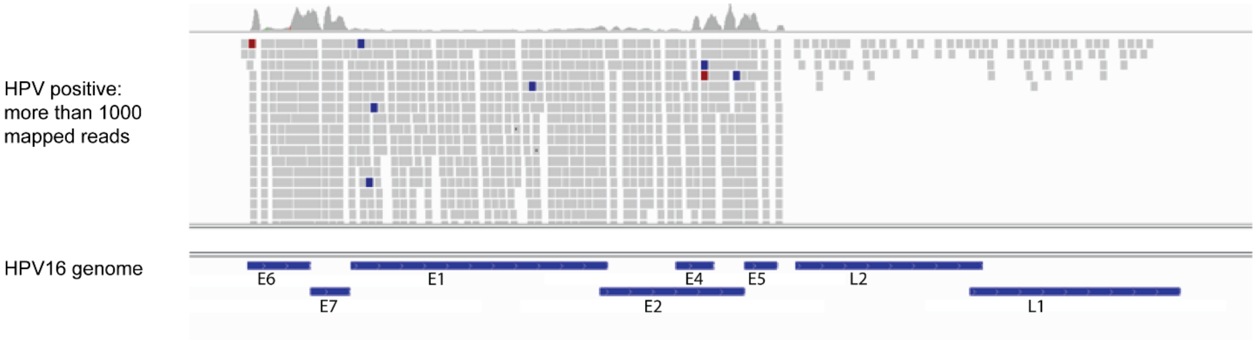


Figure S1.1, part 3. Example of RNA sequence alignments in the HPV type 16 genome. Integrated genome viewer software display (http://www.broadinstitute.org/igv/) of RNA sequencing read alignments for one HNSCC tumor against the HPV type 16 genome. The top panel illustrates a representative HPV(+) sample with more than 1000 mapped reads. A linearized version of the HPV16 genome is shown in the bottom panel.

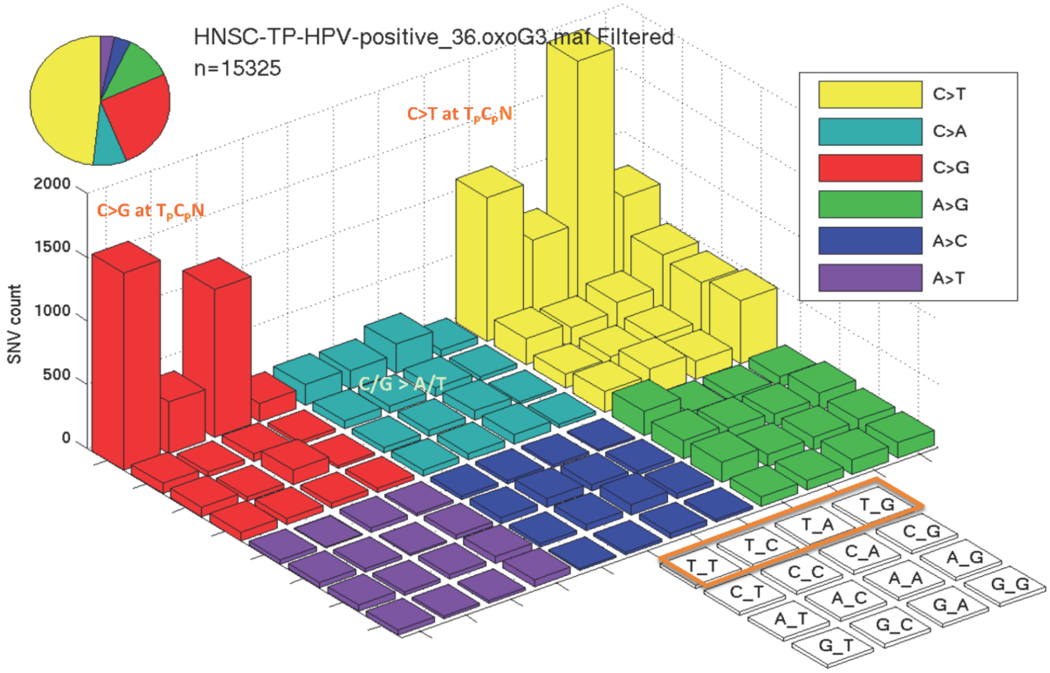


Figure S1.1, part 4. Single nucleotide variants (SNV) mutation count lego plot for HPV(+) samples. Each bin is normalized by base coverage for that bin. Colors represent the six SNV types on the upper right. The three base content of each mutation is labeled in the 4 x 4 legend on the lower right. HPV(+) samples show an enrichment for APOBEC signature mutations C > G and C > T in TpCpN trinucleotides.

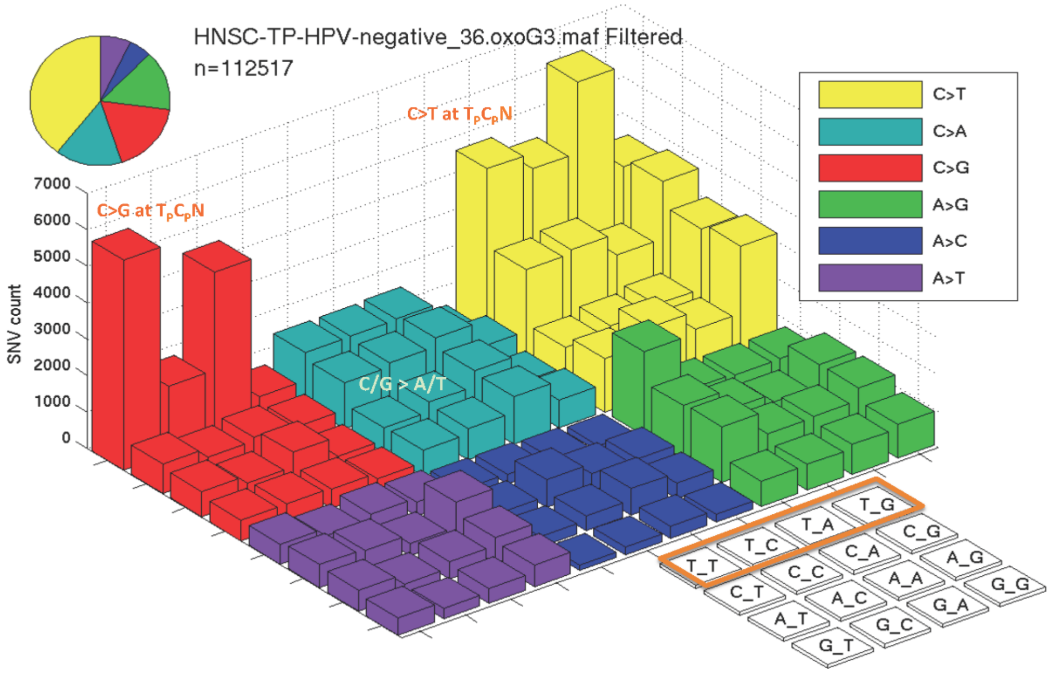


Figure S1.1, part 5. SNV mutation count lego plot for HPV(-) samples. Each bin is normalized by base coverage for that bin. Colors represent the six SNV types on the upper right. The three base content of each mutation is labeled in the 4 x 4 legend on the lower right. HPV(-) samples show an enrichment for smoking signature mutations C > A.

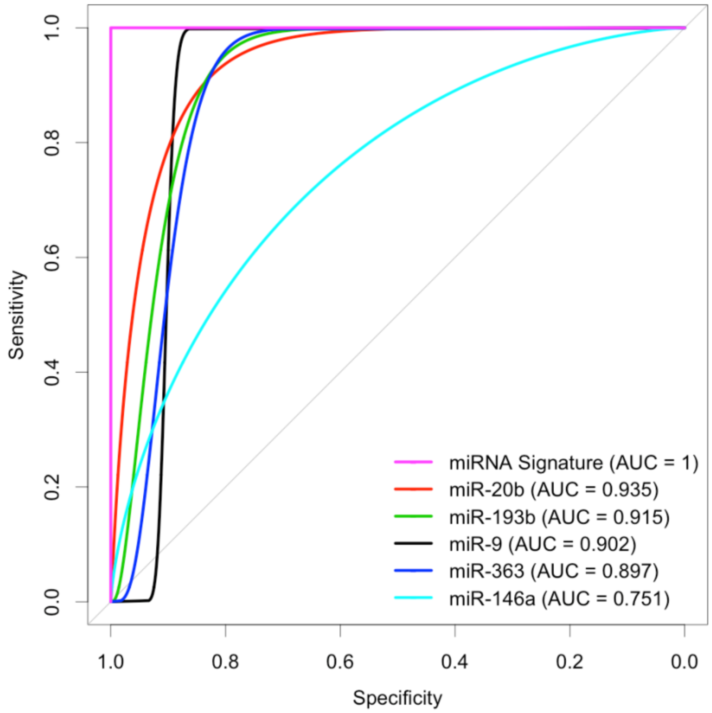


Figure S1.2. Receiver operating characteristic (ROC) curves in HPV-associated miRNAs in oropharyngeal HNSCC. Also shown are areas under the curve (AUCs) of individual miRNAs.

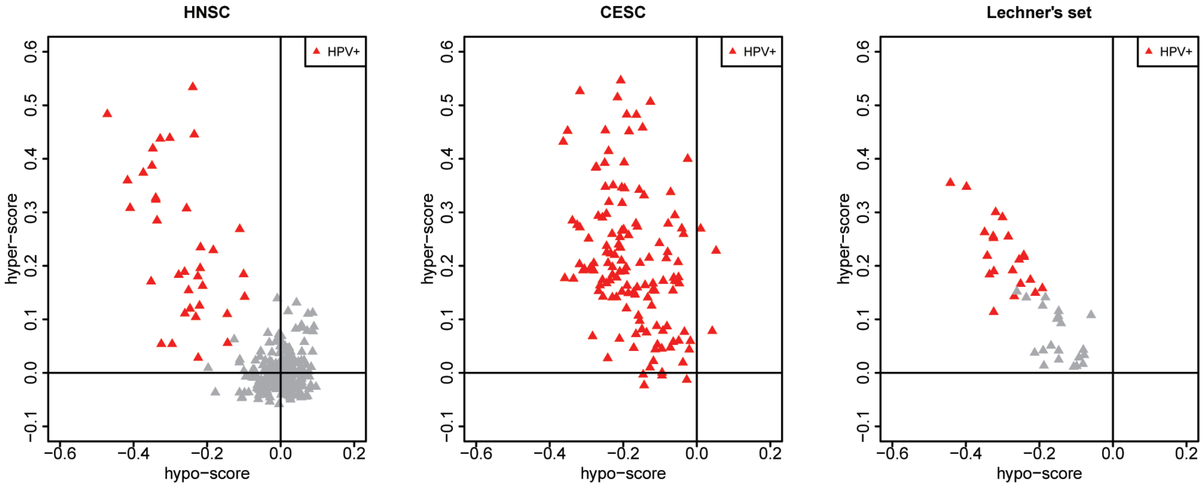


Figure S1.3. DNA methylation signatures of HPV. HPV positive samples are shown in red, Cervical squamous cell carcinoma (CESC) samples that were not explicitly evaluated are marked as HPV(+) to reflect the very strong association with cervical cancer. HNSC, head and neck squamous cell carcinoma; hypo-score, hypomethylation score; hyper-score, hypermethylation score.

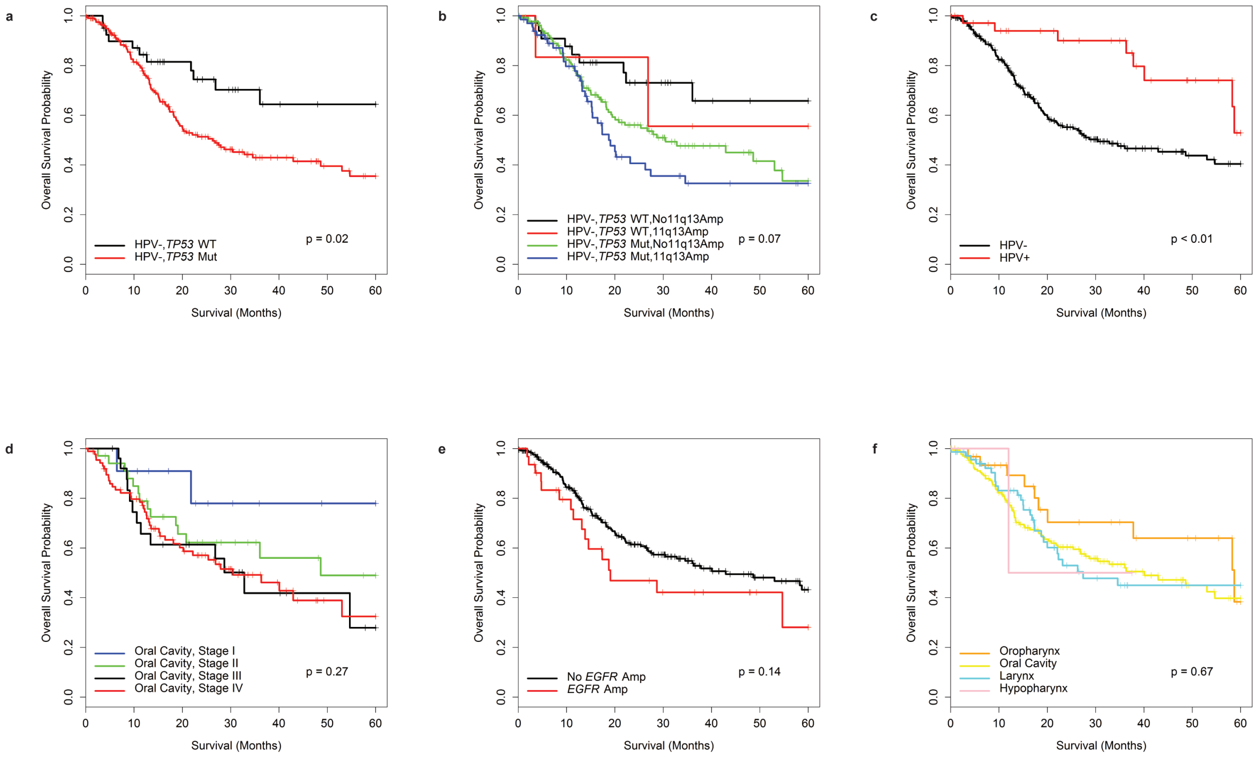


Figure S1.4. Survival analysis for select clinical and genomic variables. Kaplan-Meier plots and log rank test p-values comparing overall survival times by (a) *TP53* mutation status in HPV(-) patients, (b) *TP53* mutation/chr11q13 amplification status in HPV(-) patients, (c) HPV status in all patients, (d) tumor stage in oral cavity patients, (e) *EGFR* amplification status in all patients, and (f) tumor site in all patients. Amp, amplification.

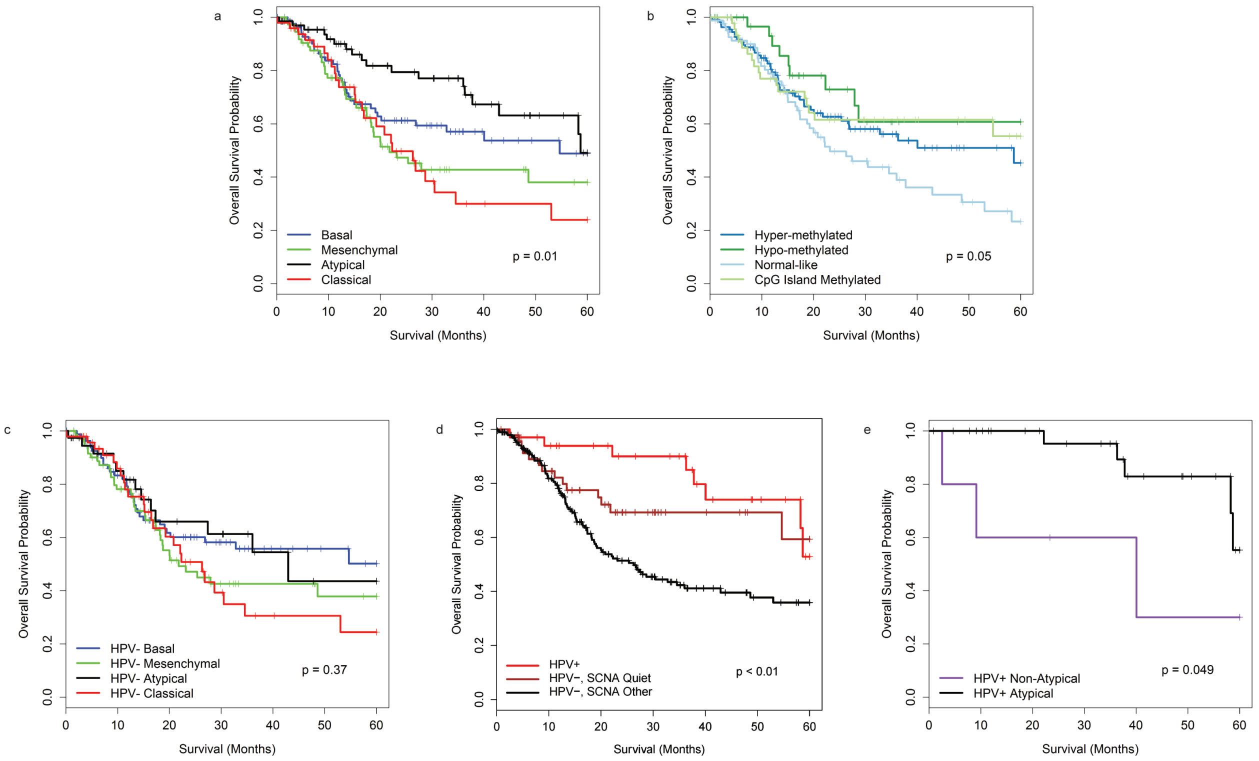


Figure S1.5. Survival analysis for platform specific subtypes. Kaplan-Meier plots and log rank test p-values comparing overall survival times among groups defined by (a) RNA subtypes for all samples, (b) methylation subtypes for all samples, (c) RNA subtypes for HPV(-) samples, (d) HPV status and DNA copy number subtypes for all samples, and (e) atypical vs. non-atypical RNA subtype for HPV(+) samples. RNA subtypes, methylation subtypes, and DNA copy number subtypes are as described in S5.1, S7.3, and S2.1, respectively. DNA copy number subtype three is denoted somatic copy number alteration (SCNA) quiet because it exhibits fewer copy number events.

Variable	Overall	HPV(-)	HPV(+)	BA	MS	AT	CL
Num. Patients	279	243	36	87	75	68	49
Age							
Age (median)	61	62	59	60	64	60	62
Age (# < 40)	14	12	2	6	3	4	1
Gender							
Female	76	72	4	26	27	14	9
Male	203	171	32	61	48	54	40
Race							
Black	25	23	2	7	5	7	6
White	242	208	34	72	68	61	41
Other	4	4	0	2	1	0	1
Race NA	8	8	0	6	1	0	1
Alcohol History							
No Alcohol Use	85	80	5	30	27	13	15
Alcohol Use	188	158	30	57	46	54	31
Alcohol NA	6	5	1	0	2	1	3
Smoking History							
Non-smoker	52	42	10	19	16	15	2
Reformed > 15	49	45	4	18	19	9	3
Reformed <= 15	81	69	12	21	19	17	24
Current smoker	90	80	10	28	18	26	18
Smoking NA	7	7	0	1	3	1	2
Mean(Packyears)	50.6	53	31.8	47.7	44.8	50.6	62
Site							
Oropharynx	33	11	22	4	6	22	1
Oral Cavity	172	160	12	76	56	20	20
Larynx	72	71	1	6	13	25	28
Hypopharynx	2	1	1	1	0	1	0
Stage							
Stage I	14	12	2	7	4	3	0
Stage II	44	39	5	16	13	8	7
Stage III	38	34	4	11	7	11	9
Stage IVa	139	127	12	42	43	27	27
Stage IVb	5	5	0	2	1	1	1
Stage NA	39	26	13	9	7	18	5
Tumor Status							
T0-T2	92	74	18	28	25	24	15
T3-T4	161	143	18	48	44	40	29
Tumor Status NA	26	26	0	11	6	4	5
Node Status							
N0-N2	183	165	18	58	50	40	35
Above N2	70	52	18	18	19	24	9
Node Status NA	26	26	0	11	6	4	5
HPV Status							

HPV(-)	243	243	0	84	73	38	48
HPV(+)	36	0	36	3	2	30	1

Table S1.1, part 1. Summary of clinical data. Data are presented as a function of all cases (N=279), by HPV status, and as a function of the gene expression subtypes basal (BA), mesenchymal (MS), atypical, (AT), and classical (CL).

Variable	Overall	HPV(-)	HPV(+)	Larynx	Oral Cavity	Oropharynx
Num. Patients	279	243	36	74	172	33
Age						
Age (median)	61	62	59	61	62.5	56
Age (# < 40)	14	12	2	2	9	3
Gender						
Female	76	72	4	14	56	6
Male	203	171	32	60	116	27
Race						
Black	25	23	2	12	10	3
White	242	208	34	59	153	30
Other	4	4	0	1	3	0
Race NA	8	8	0	2	6	0
Alcohol History						
No Alcohol Use	85	80	5	24	59	2
Alcohol Use	188	158	30	48	110	30
Alcohol NA	6	5	1	2	3	1
Smoking History						
Non-smoker	52	42	10	3	39	10
Reformed > 15	49	45	4	8	37	4
Reformed <= 15	81	69	12	25	46	10
Current smoker	90	80	10	36	45	9
Smoking NA	7	7	0	2	5	0
Mean(Packyears)	50.6	53	31.8	60.8	46.1	43.8
Site						
Oropharynx	33	11	22	0	0	33
Oral Cavity	172	160	12	0	172	0
Larynx	72	71	1	72	0	0
Hypopharynx	2	1	1	2	0	0
Stage						
Stage I	14	12	2	0	11	3
Stage II	44	39	5	6	34	4

Stage III	38	34	4	9	26	3
Stage IVa	139	127	12	45	88	6
Stage IVb	5	5	0	2	2	1
Stage NA	39	26	13	12	11	16
Tumor Status						
T0-T2	92	74	18	13	62	17
T3-T4	161	143	18	54	92	15
Tumor Status NA	26	26	0	7	18	1
Node Status						
N0-N2	183	165	18	49	121	13
Above N2	70	52	18	18	33	19
Node Status NA	26	26	0	7	18	1
HPV Status						
HPV-	243	243	0	72	160	11
HPV+	36	0	36	2	12	22

Table S1.1, part 2. Summary of clinical data. Data are presented as a function of all cases (N=279), by HPV status, and as a function of tumor site. NA, not available.

Variable	Overall	HPV(-)	HPV(+)	CN 1	CN 2	CN3/Quiet
Num. Patients	279	243	36	109	98	72
Age						
Age (median)	61	62	59	60	61	64
Age (# < 40)	14	12	2	5	5	4
Gender						
Female	76	72	4	28	25	23
Male	203	171	32	81	73	49
Race						
Black	25	23	2	12	10	3
White	242	208	34	93	82	67
Other	4	4	0	2	2	0
Race NA	8	8	0	2	4	2
Alcohol History						
No Alcohol Use	85	80	5	31	28	26
Alcohol Use	188	158	30	75	67	46
Alcohol NA	6	5	1	3	3	0
Smoking History						
Non-smoker	52	42	10	22	13	17
Reformed > 15	49	45	4	13	13	23
Reformed <= 15	81	69	12	35	30	16
Current smoker	90	80	10	35	40	15
Smoking NA	7	7	0	4	2	1
Mean(Packyears)	50.6	53	31.8	50.8	61.6	34.8
Site						
Oropharynx	33	11	22	10	6	17
Oral Cavity	172	160	12	70	53	49
Larynx	72	71	1	28	39	5
Hypopharynx	2	1	1	1	0	1
Stage						
Stage I	14	12	2	6	4	4
Stage II	44	39	5	13	15	16
Stage III	38	34	4	10	15	13
Stage IVa	139	127	12	66	46	27
Stage IVb	5	5	0	1	4	0
Stage NA	39	26	13	13	14	12
Tumor Status						
T0-T2	92	74	18	31	25	36
T3-T4	161	143	18	64	63	34
Tumor Status NA	26	26	0	14	10	2
Node Status						
N0-N2	183	165	18	73	60	50
Above N2	70	52	18	22	28	20
Node Status NA	26	26	0	14	10	2
HPV Status						

HPV-	243	243	0	98	98	47
HPV+	36	0	36	11	0	25

Table S1.1, part 3. Summary of clinical data. Data are presented as a function of all cases (N=279), by HPV status, and as a function of DNA copy number subtype, copy number subtype 1, CN1; copy number subtype 2, CN2; and copy subtype number 3, quiet, CN3/quiet

S2: Copy number analysis

S2.1 SNP array -based copy number analysis

DNA from each tumor or germline-derived sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described[24]. Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy number at each probe locus[25]. For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor [26]. This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy number profile. Individual copy number estimates then underwent segmentation using Circular Binary Segmentation [27]. As part of this process of copy number assessment and segmentation, regions corresponding to germline copy number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy number changes underlying each segmented copy number profile [28]. Analysis of broad copy number alterations was then conducted as previously described[28]. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0 [28]. Statistically significant regions as well as genes encompassed for the regions are provided for HNSCC (all samples and by HPV status), LUSC, CSCC are shown in Figures S2.1-S2.3 and provided as Data File S2.1. Thresholded copy number at reoccurring alteration peaks from GISTIC 2.0 analysis of the entire set (all_lesions.conf_99.txt file from the GISTIC output) was used for copy number based clustering and comparison of focal peak frequencies across HPV status and tumor location. Tumors were hierarchically clustered in R based on Euclidean distance using Ward's method. Fisher exact p values were calculated for frequency comparisons of significantly reoccurring alterations by HPV status and site, as shown in Data File S2.2. For comparison of amplifications only high-level events were considered (thresholded values = 2); for deletions all events were used. Allelic copy number, and purity and ploidy estimates were calculated using the ABSOLUTE algorithm [29]. Additional results from all HNSCC copy number analyses can be found at http://gdac.broadinstitute.org/runs/awg_hnsc__2014_04_12/reports/. TCGA copy number data from Lung Squamous Cell Carcinoma (LUSC) and Cervical Squamous Cell Carcinoma (CSCC) are from the February 22nd 2013 Broad Firehose analysis (Figure S2.2) [30].

S2.2 Structural alterations

Structural alterations including copy number changes and loss of heterozygosity were investigated using the data described above. Structural rearrangements were assessed primarily through RNA-Seq analysis of fusion transcripts (Section S3, "high coverage" whole genome sequencing (30X, n=29), and "low pass" whole genome sequencing (5X, n=98) (Section S4). We focused on alterations with potential clinical implication, especially driver genes, and genome integration of HPV in HNSCC.

Additional Findings

Most tumors demonstrated copy number alterations (CNAs) including losses of 3p and 8p, and gains of 3q, 5p, and 8q chromosomal regions (Figures 1A and S2.1, Section S2) resembling lung squamous cell carcinomas (LUSCs) [31] (Figure 1A), underscoring shared molecular pathogenesis (Figures S2.1, S2.2). Overall levels of genomic instability were high, as quantified by the number of copy number segments. As shown in Figure S2.4, HPV(-) subjects have more copy number segments (median = 136) than HPV(+) subjects (median = 113), one-sided Mann-Whitney p-value = 0.026, suggesting higher levels of genomic instability in HPV(-) patients.

Unsupervised clustering identified three copy number subtypes, one of which (class three) is characterized by having fewer copy number alterations. Copy number class three is enriched for "M class" tumors that are driven primarily by mutations, not copy number alterations (Figure 1) [32]. Although more M class patients were HPV(-) vs HPV(+), (65/86, 76% vs 21/86, 24%, respectively), a higher proportion of HPV(+) (21/36, 58%) was M class relative to HPV(-) (65/243, 27%).

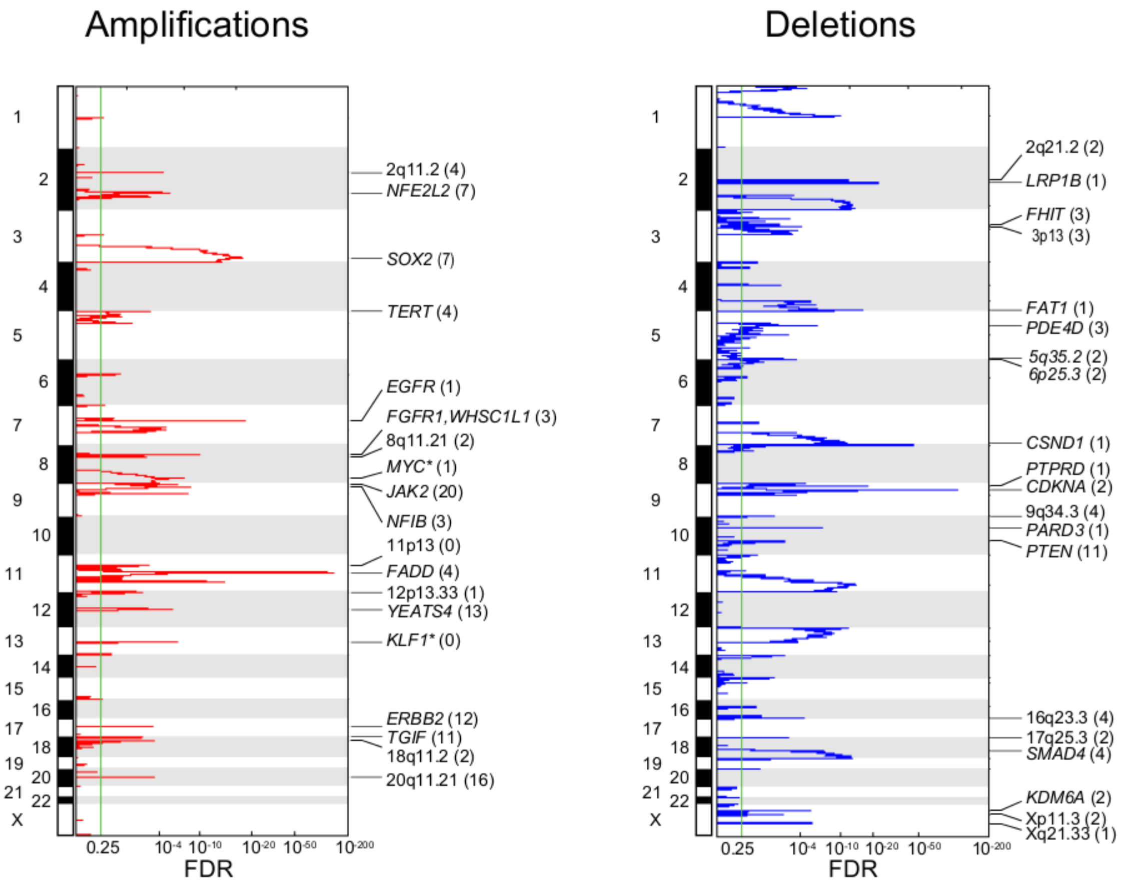


Figure S2.1. GISTIC 2.0 analysis of significantly reoccurring focal alteration in 279 head and neck squamous cell tumors. Regions of significantly reoccurring amplifications (left) and deletions (right) are plotted by false discovery rate. Annotated peaks have residual q values less than 0.5 and 20 or fewer genes within peak regions. These peak regions are annotated with possible driver genes and the total number genes within these peaks. Genes marked with * are those not located within regions defined by GISTIC, but are altered in the majority of samples within a peak. Details on the segments and genes are provided as Data File S2.1. FDR, false discovery rate.

Amplifications

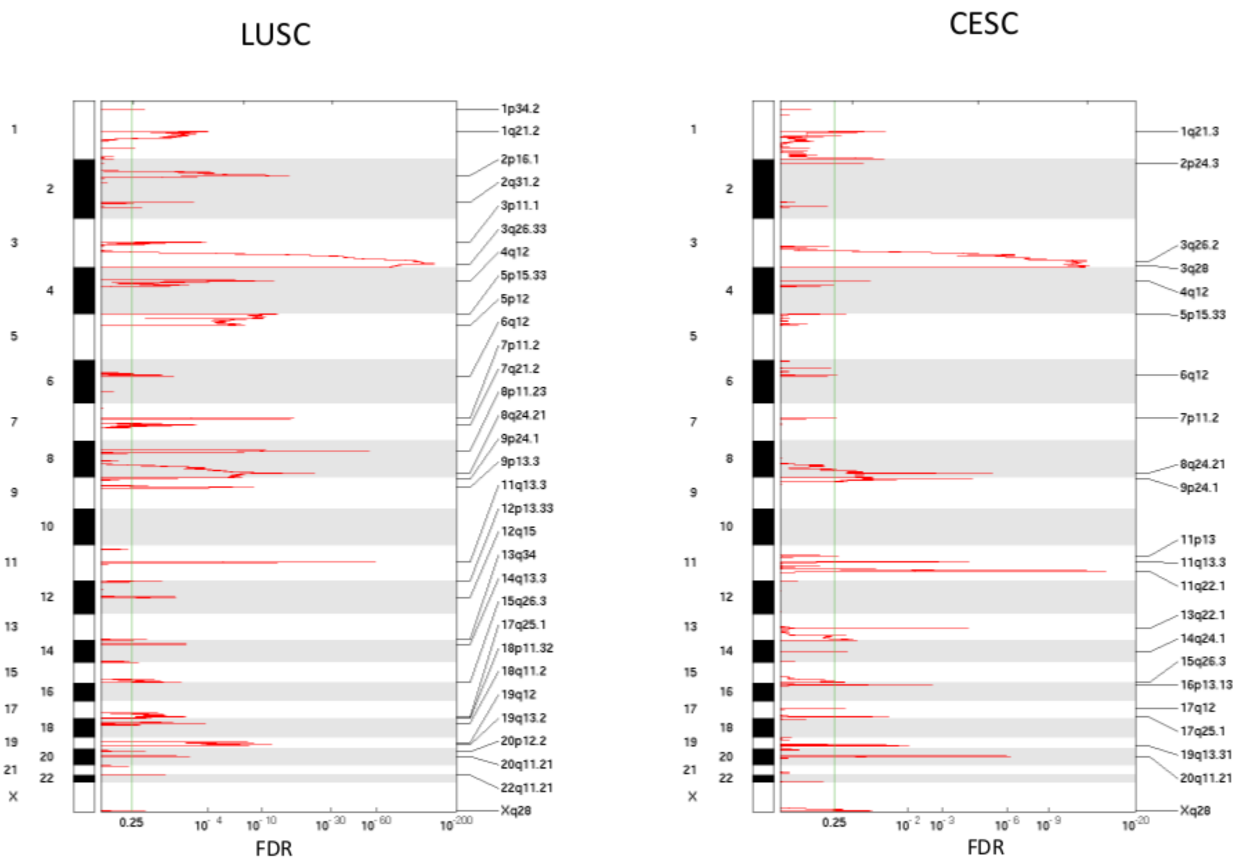


Figure S2.2, part 1. GISTIC amplification peaks in lung squamous cell carcinoma (LUSC) and cervical squamous cell carcinoma (CESC). Details on the segments and genes are provided as Data File S2.1. FDR, false discovery rate.

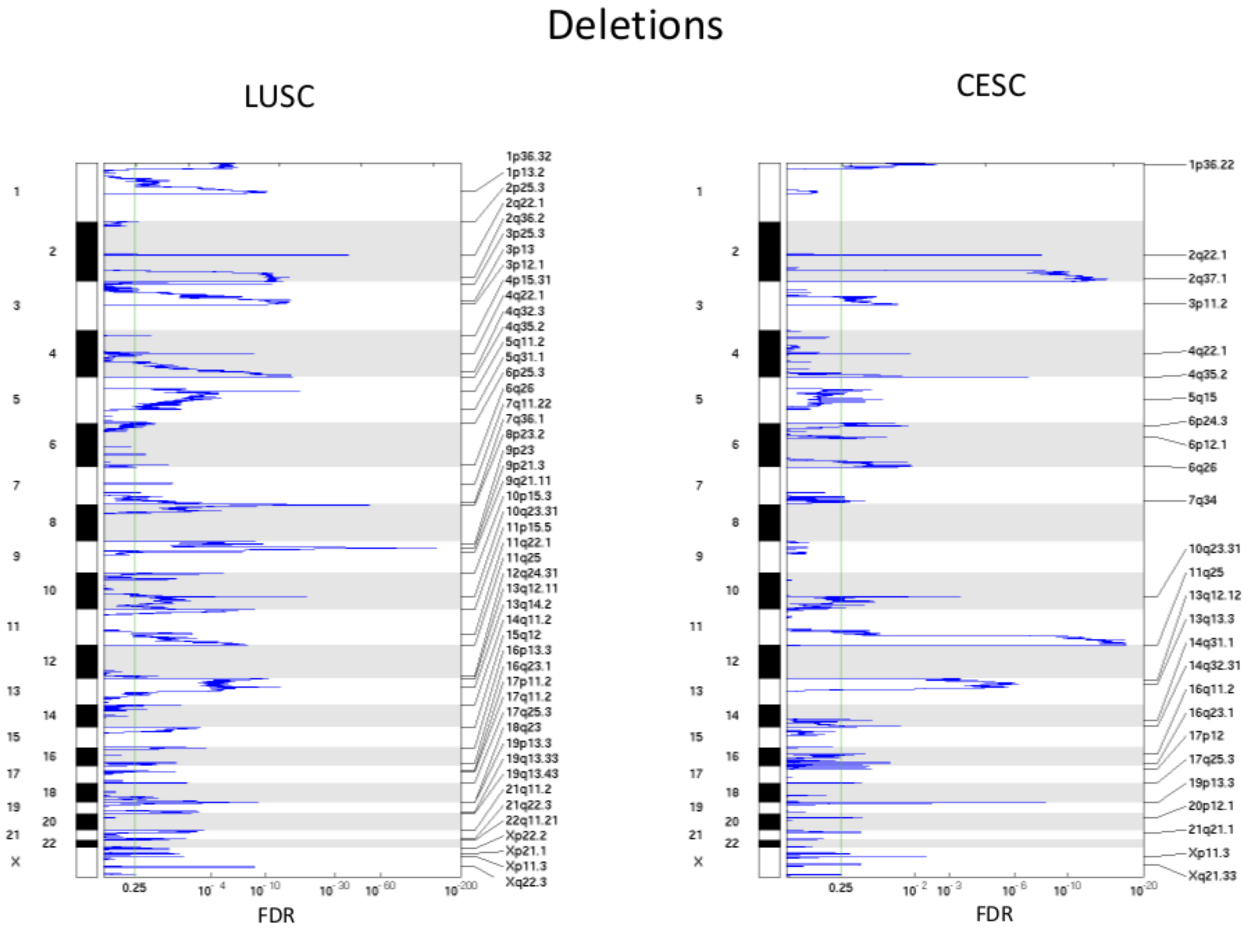


Figure S2.2, part 2. GISTIC deletion peaks in lung squamous cell carcinoma (LUSC) and cervical squamous cell carcinoma (CESC). Details on the segments and genes are provided as Data File S2.1. FDR, false discovery rate.

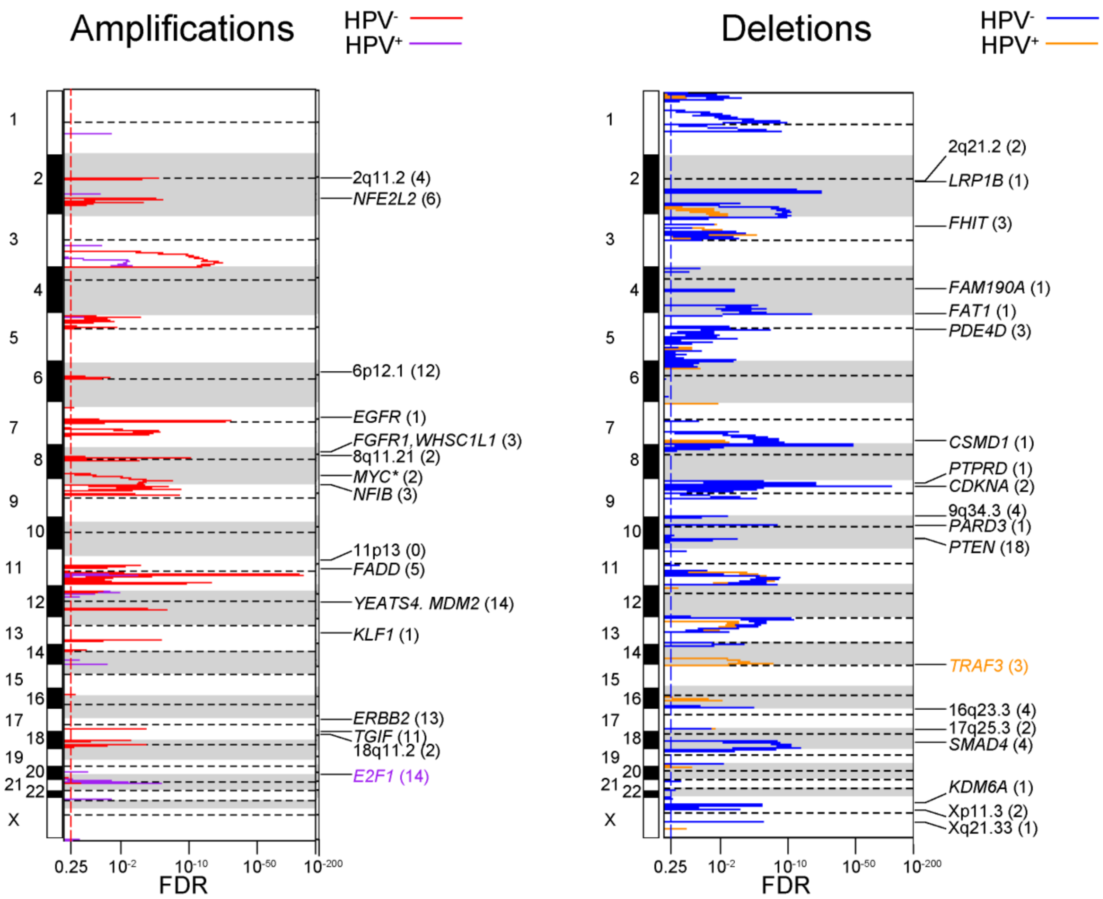


Figure S2.3. Comparison of GISTIC 2.0 analyses of 243 HPV(-) and 36 HPV(+) head and neck tumors. Regions of significant amplifications (left) and deletions (right) are plotted by false discovery rate. GISTIC plots from HPV(+) tumors (purple and orange lines) are superimposed onto GISTIC plots from HPV(-) tumors (red and blue lines). Peak regions annotated in black are those from HPV(-) GISTIC 2.0 analysis. Annotated HPV(-) peaks have residue q values less than 0.5 and 20 or fewer genes within peak regions. Peaks are shown with possible driver genes and the total number of genes within these peaks. Genes marked with * are those not located within regions defined by GISTIC 2.0, but are altered in the majority of samples within a peak. Also annotated are the HPV(+) specific amplification peak containing *E2F1* and deletion peak containing *TRAF3*. Details on the segments and genes are provided as Data File S2.1. FDR, false discovery rate.

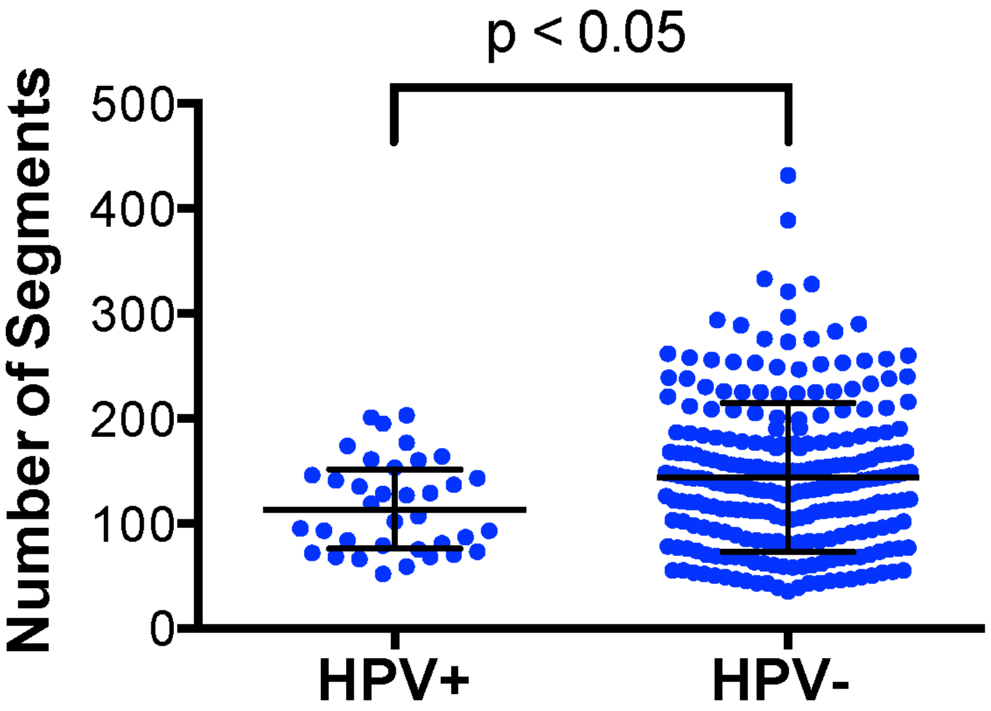


Figure S2.4. Number of copy number segments in HPV(+) and HPV(-) samples. Copy number segments for each subject were segmented using Circular Binary Segmentation [27], as described in supplementary text above. The number of copy number segments for each subject is shown according to their HPV status. Horizontal bars show the median and interquartile range in each group of subjects.

S3. RNA sequencing

S3.1 RNA sequencing and expression quantification

Methods for sequencing and data processing of RNA using the RNA-Seq protocol have been previously described for TCGA[31]. Briefly, RNA was extracted, prepared into Illumina TruSeq mRNA libraries, and sequenced by Illumina HiSeq2000 with a target of 60 million read-pairs per tumor resulting in paired 48nt reads, and subjected to quality control as previously described[31]. RNA reads were aligned to the hg19 genome assembly using Mapsplice [6]. Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1 [33] using RSEM [34] and normalized within-sample to a fixed upper quartile of total reads. For further details on this processing, refer to the Description file at the DCC data portal under the V2_MapSpliceRSEM workflow [35,36]. For gene level analyses, expression values of zero were set to the overall minimum value, and all data were log2 transformed [37].

S3.2 RNA-Seq for confirmation of somatic alterations reported in whole exome sequencing

For the purposes of confirmation of somatic alterations, we limited the analysis to somatic positions reported in the DNA WES mutation annotation format file (MAF). WES methods are described in Section S4. Somatic variants reported in the HNSCC MAF associated with the data freeze [37] were interrogated for confirmation in RNA-Seq using the tool UNCeqR as previously described [31,38] (http://lbg.med.unc.edu/tools/unceqr/). Briefly, WES mutation positions having a minimum of 1X RNA depth were evaluated for the presence of at least one read confirming the variant allele (https://tcga-data.nci.nih.gov/docs/publications/hnsc_2014/mafX.hnscc.20140723.csv). Under this condition, 71% of mutation loci were expressed, 50% were confirmed, and 70% were confirmed if the locus was expressed (at 1X), on average across samples (Figure S3.1a). When more stringent coverage requirements are considered, the confirmation rates increase while positions covered decreases. Increasing the minimum to 5X of RNA coverage depth before considering the DNA variant position evaluable while still requiring at least one variant read in the RNA, only 58% of mutation alleles were expressed (Figure S3.1b). However, the confirmation rate increases to 86% in covered positions (46% of all positions when including positions not considered covered).

S3.3 Gene fusion detection

In addition to quantifying gene expression and detecting mutations, RNA-Seq can detect structural variants including alternate splicing, intra-chromosomal fusions, and inter-chromosomal fusions. For the purposes of the current analysis we relied on the MapSplice implementation of fusion detection. Briefly, any two segments of a read alignment that (1) are separated by a gap longer than 300,000 nt, or (2) are on different chromosomes, or (3) are on different strands, or (4) map to disordered locations (i.e. the apparent direction of transcription changes between the segments) were identified as fusion candidates. In order to decrease false positives, these candidates were further filtered by (1) requiring surrounding read alignments to also exhibit the fusion, and (2) requiring 25bp of donor and acceptor sequences to have unique alignment on the genome (Data File S3.1), and in a subset of manually curated cases (3) visually examining predicted fusion events of special interest utilizing a novel realignment and visualization utility distributed with MapSplice version 2.0.1.9. For each predicted fusion, this visualization tool generates a contiguous synthetic genomic reference sequence across the fusion junction. This region includes the sequence from both the donor and acceptor sides of a putative fused transcript plus flanking genome sequence immediately adjacent to the predicted genomic fusion loci. An attempt is then made to (re)align all reads from the RNA-Seq experiment that predicted the fusion to the synthetic fusion reference sequences. All the reads that map to one of the synthetic fusion loci (including flanking regions) are collected into one BAM file, those reads that support the fusion are also copied into a second more exclusive BAM file. This second file contains only reads directly supporting the fusion junction, either by spanning it or comprising a mate pair that bridges the junction even though neither read spans it. These BAM files together with the synthetic fusion sequences can be loaded into conventional software such as the Integrative Genomics Viewer (IGV)[39] for the purposes of visualizing the predicted fusion events as well as its read alignments. Visualization of predicted fusions in this way provides an opportunity for the application of human pattern recognition skills to the task of filtering fusions through direct qualitative inspection of the predicted variant and its supporting reads—bridging and spanning—within the context of its surrounding genomic sequence and transcript models (Figure S3.2).

Using the automated components of the filtering process described above, MapSplice generated a set of 13,759 predicted fusion events in 279 samples with a median of 45 fusion events per sample (Data File S3.1). The predicted fusions were then clustered by genomic position of both the donor and acceptor site in order to compare closely related fusions across samples and collapse multiple fusions in close proximity in a single sample, which might be observed due to alternate splicing around a fusion event. This reduced the number of fusions to 1,554 cluster loci, which corresponds to an average of 8.9 predicted fusions per cluster. Closer inspection revealed that the distribution of predicted fusions per cluster was highly skewed with the 20 largest clusters harboring a total of 7,805 fusions (56.7%). Manual review of these large clusters revealed that they typically contain pairs of genes with high sequence homology. As a result, even within a single sample many spurious fusions were predicted in close proximity on both the donor and acceptor ends of the fusion. Of the remaining 1,534 clusters, 876 were comprised of a single predicted fusion. Of the 1,534 clusters, the vast majority were represented by only a few bridging reads, suggesting overall low expression levels of the fusion transcripts. Eighty percent of all predicted fusion events were intra-chromosomal (as opposed to inter-chromosomal), and fewer than 20% of predicted fusions were determined to be in-frame. Taken together, the data suggest that most predicted fusions result in low levels of transcription across break points within a chromosome, perhaps because of copy number alterations that produce truncated genes.

S3.4 RNA-Seq for gene splicing and viral integration

In S3.3 we defined a fusion as any two segments of a read alignment that (1) are separated by a gap longer than 300,000 nt, or (2) are on different chromosomes, or (3) are on different strands or (4) map to disordered locations (i.e. the apparent direction of transcription changes between the segments). By definition, any two segments of a read alignment that are separated by a gap less than 300,000 nt would be considered as an RNA splice rather than a fusion. Splicing information for each sample is available through DCC (see files named "junction_quantification.txt" at https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/hnsc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_HNSC.IlluminaHiSeq_RNASeqV2.Level_3.1.6.0/), whereas putative fusions are available in Data File S3.1.

The definition implies that when considering the RNA data alone, it is not possible to differentiate with certainty among different etiologies of structural variants. Specifically, it is possible that some spliced alignments defined as splices might actually be the result of DNA deletions less that 300,000 nt. From experience this distinction is important when considering loss of function events in tumor suppressor genes which may be found either in the spliced alignments primarily or in the file containing predicted fusions.

We undertook an integrated manual review of these data to investigate genes with known or suspected oncogenic splicing, most notably *EGFR* (Figure S3.3) and *MET* (Figure S3.4), and identified a limited number of cases. More commonly, altered splicing resulted in loss of function in known tumor suppressors such as was previously described in the TCGA lung squamous cell carcinoma project for the gene *CDKN2A* and its protein coding transcript p16INK4A and p14ARF [31] (Figures S3.5 and S3.6) documenting multiple mechanisms for gene disruption. We extended the approach to the novel tumor suppressor *FAT1* (Figures S3.7 and S3.8) and to proposed driver genes in HNSCC (Figure S3.9). Taken together, these data suggest that there are diverse mechanisms for inactivation of tumor suppressors in HNSCC and that integrating data from multiple platforms is important for identifying all tumor suppressor events, which would have been underestimated if sequence mutations alone were considered.

We then interrogated HPV viral integration sites defined in the “HPV detection by multiple nucleic acids techniques” section of S1.2 to determine the potential impact on transcription (Figure S3.10, Data File S3.2) with a striking finding that almost every case of viral integration occurred either within or near a protein coding transcript. The functional impact, if any, of integration events is speculative. Given that none of the events were recurrent and that some occurred in gene introns, loss of gene function might be predicted. Arguing against loss of function is the fact that most integration sites occurred in the setting of copy number amplifications. More functional data are required to make further comment on the impact of specific integration sites.

Methods have been proposed to detect alternative spicing across samples assayed by RNA-Seq. We selected the SigFuge[40] approach used in prior TCGA reports[31] which selects genes in which at least two isoforms are expressed differentially across a sample set, such that some samples express isoform A while others express isoform B preferentially. SigFuge was applied to the 279 samples across a set of 20,500 genes defined by TCGA general annotation file v2.13. Using a FDR cutoff of 0.05, 335 genes were identified as having significantly differential isoform usage (Data File S3.3). SigFuge uses genome coordinates rather than transcript coordinates to identify regions of differential gene expression within genes. This approach is documented to be more powerful for detecting regions of differential expression, but introduces the non-trivial challenge of resolving novel and existing transcript definitions. Given the scope of this project, we provide the named genes in which alternate transcripts were detected by this method but leave a more detailed analysis for future work.

Two examples from the SigFuge analysis are shown in the supplemental figures: 1) a novel example of *KLK12* alternate transcription (Figure S3.11), 2) an expected example of the well documented alternative transcription of *TP63* (Figure S3.12). *KLK12* is part of a family of kallikrein-related peptidases (KLK genes) with splice variants receiving increased attention as potential biomarkers in cancer[41]. The three panels in Figure S3.11 show the log-transformed per-base read depth for the 279 samples. Each curve corresponds to the expression for a single sample. Cluster 1 consists of samples with low overall expression, and Clusters 2 and 3 differ by their usage of exon 4. Expression of exon 4 is clearly present in samples included in Cluster 2 and absent from samples in Cluster 3. This directly corresponds to preferential usage of known alternate *KLK12* isoforms differing by cassette exon 4. In addition to the unsupervised approach, we used SigFuge to perform a supervised evaluation of a select set of known alternative isoform usage or splicing events for genes that play a role in HNSCC, including *TP63*. Coverage and splicing was evaluated at relevant exons and splice sites, and for *TP63* this identified a predominance of the DeltaN isoform of the gene, as expected (Figure S3.12).

However, as panels A and B of the figure show, there was heterogeneity in the isoform usage across samples, the full functional implications of which are unknown. In addition, multiple samples showed evidence of predicted fusions involving *TP63*. While none of the fusions were recurrent, the example shown in panel C produced a transcript which includes the DNA binding domain of the DeltaN isoform. In short, these examples illustrate that analysis of RNA sequencing data identifies events involving alternative splicing and structural alterations and thereby captures additional heterogeneity in key cancer genes that would not be easy to detect using other genomic platforms.

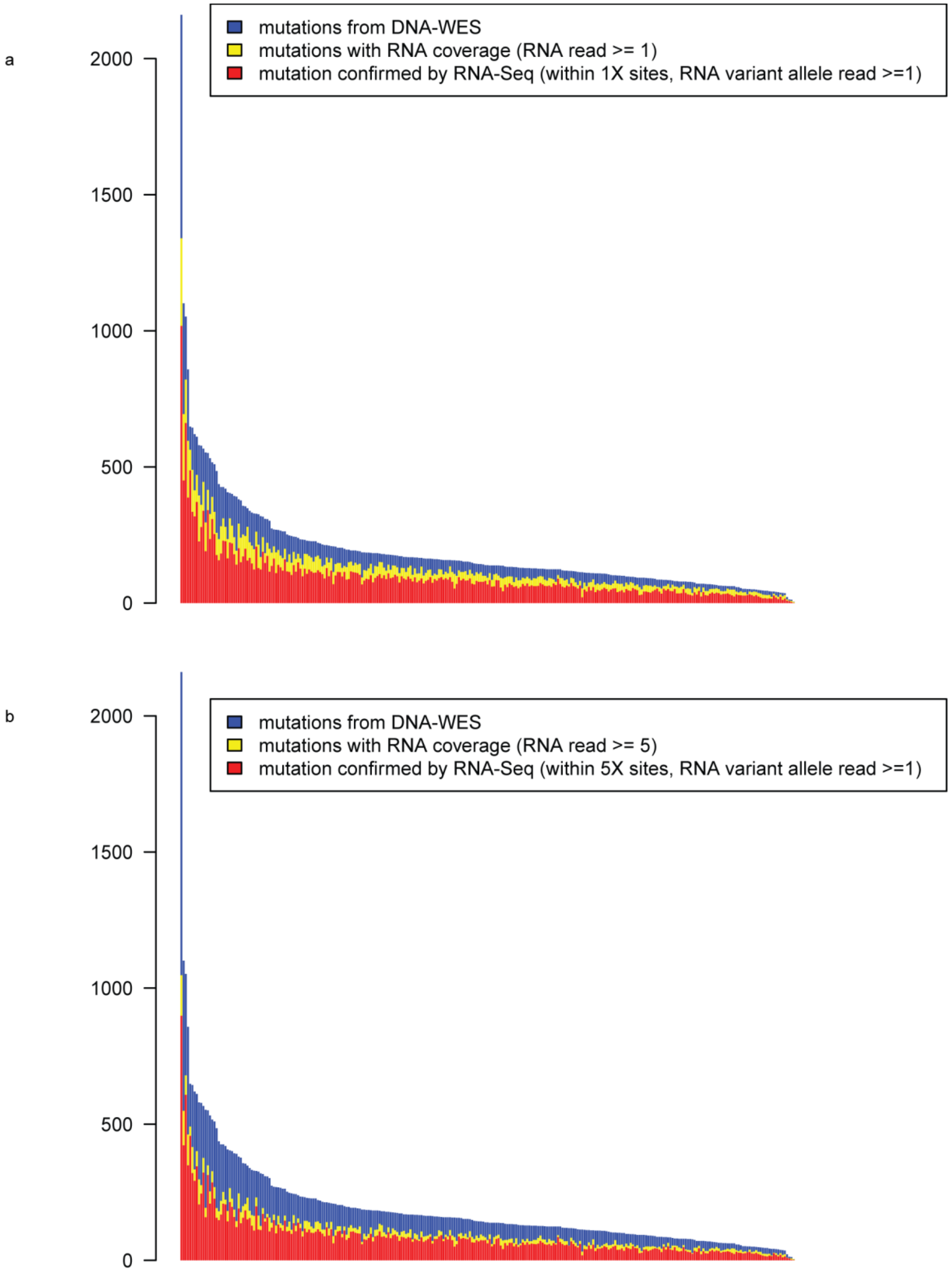


Figure S3.1. RNA-Seq for confirmation of somatic alterations reported in whole exome sequencing. Panel (a) considers positions with only a single base of RNA covering the variant reported by DNA. Panel (b) considers positions with 5X RNA coverage of the variant reported by DNA.

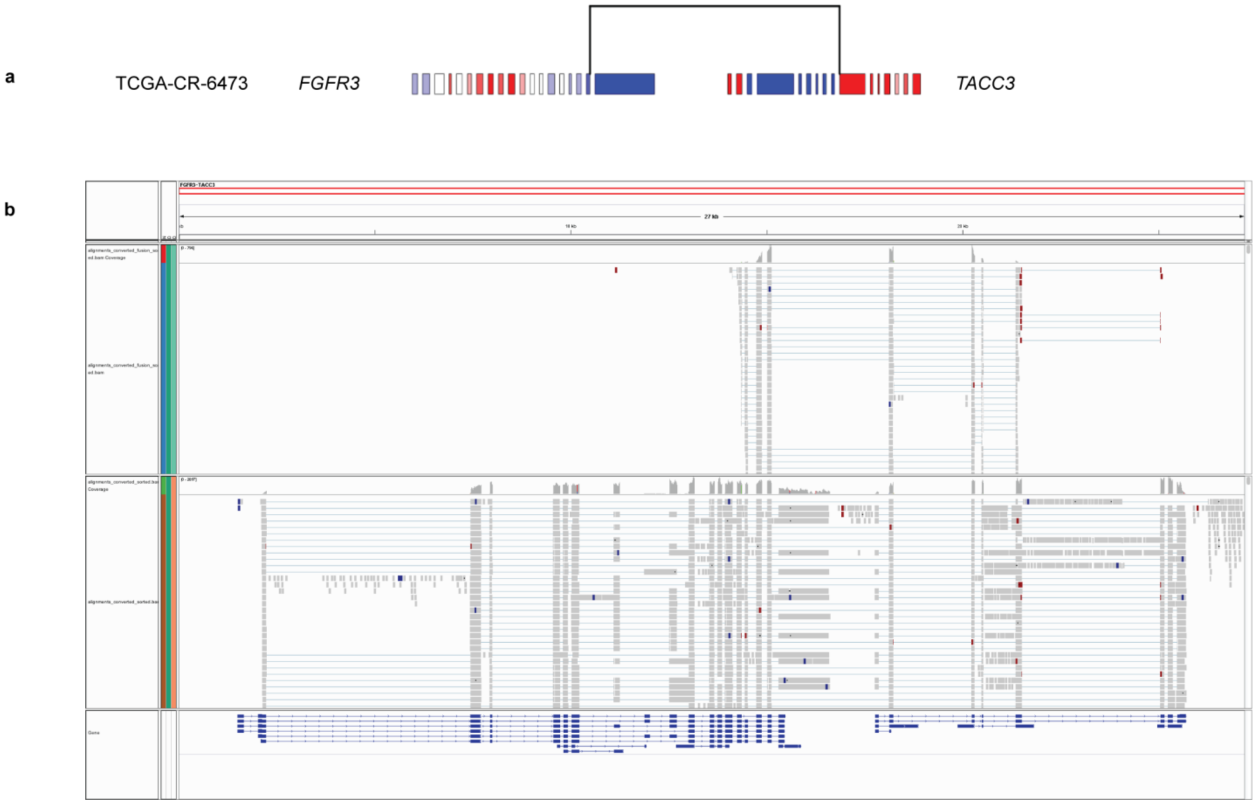


Figure S3.2. *FGFR3-TACC3* fusion event. (a) Normalized RNA expression for each exon across fusion gene partners. Red boxes indicate exons of relatively high expression compared to blue boxes indicating exons removed as a consequence of the fusion. The line indicates the predicted fusion locus and joins the two genes at that point. (b) RNA read alignments indicating *FGFR3-TACC3* fusions, which previously were reported in HNSCC [42]. The top read track shows selected reads supporting the existence of a fusion with components of the read mapping into both genes. Both spanning and bridging reads are shown. Spanning reads have contiguous sequence across the fusion junction, bridging reads have mate pairs that split across the junction, one mapping to each gene independently. The middle track displays additional reads that are not directly involved in the fusion but map to the surrounding region. With high coverage, this track clearly shows the adjacent transcript structures. At the top of each read track is a histogram representing read depth (relative but un-normalized expression level). Aligned reads are gray, with reads spanning a junction (splice or fusion) split into pieces separated by a light blue line indicating the gap between consecutively aligned bases of a read. The bottom track (dark blue) shows transcript gene models. Both *FGFR3* and *TACC3* are on the plus strand with gene models reading left to right. The *FGFR3* model is shown in full on the left, *TACC3* is shown truncated, starting at approximately exon 9. The fusion excludes the last exon of *FGFR3* and the first half of *TACC3*, with a splice junction between CHR4-1,808.661-C in *FGFR3* and CHR4-1,739.325-G in *TACC3*. This preserves the reading frame in phase across the junction and is supported by around 500 spanning or bridging reads. A drop in the expression histogram in line with the reads that span the splice to *TACC3* can be observed in the lower track across the last intron of *FGFR3*; this intron has a read depth of 1,430 on its left and 771 on its right.

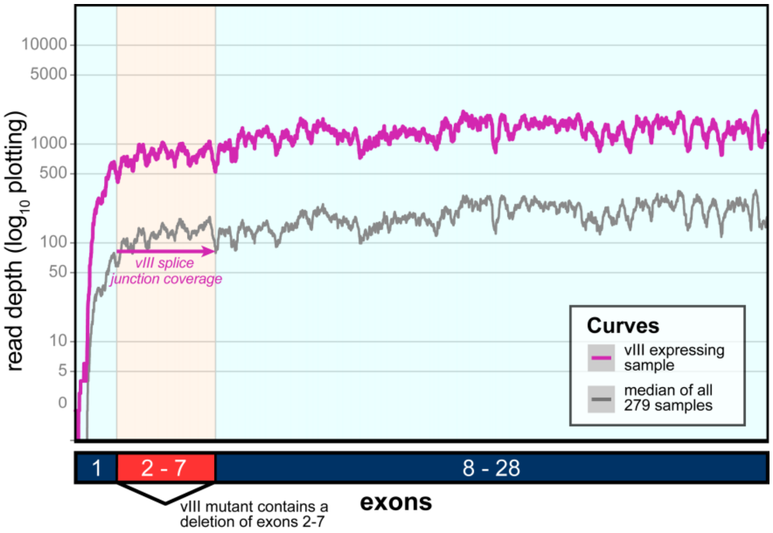


Figure S3.3. *EGFR* vIII mutant sample. A single sample (TCGA-D6-6826) was identified possessing the *EGFR* vIII mutant splice skipping exons 2-7. The per-base expressions along the exons are plotted for the single mutant sample (magenta) and the median of all 279 samples (gray). The coverage of the splice junction skipping exons 2-7 (74 reads) is denoted by the height of the magenta arrow drawn across the base positions corresponding to exons 2-7. A simple gene model is given at the bottom of the figure for reference.

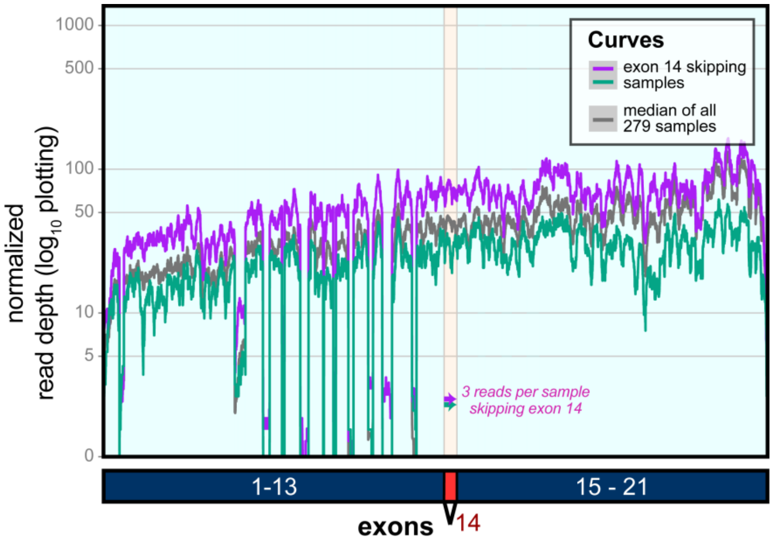


Figure S3.4. Exon 14 skipping in *MET*. Two samples (TCGA-CN-6997, TCGA-CQ-5324) were identified with evidence of exon 14 skipping. The normalized per-base expressions along the exons are plotted for the two mutant samples (purple and teal) and the median of all 279 samples. The coverage of the splice junction skipping exon 14 is denoted by the height of the purple and teal arrows. Both samples contained 3 reads spanning from exon 13 to exon 15. A simple gene model is given at the bottom of the figure for reference. Low coverage for some regions in exons 1-13 in all samples is due to inconsistency between the predicted gene model and the true transcript of the gene

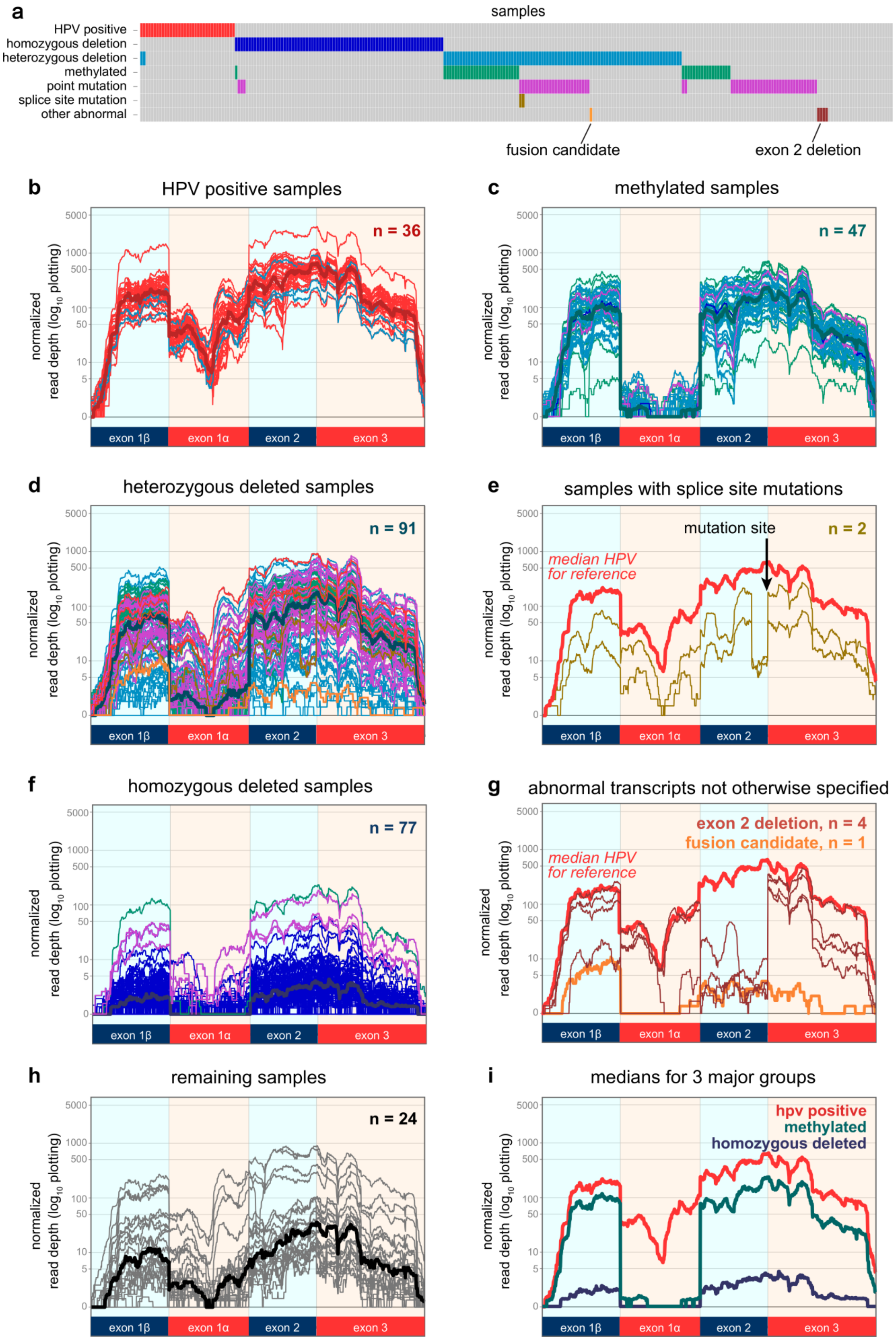

a
samples
HPV positive
homozygous deletion
heterozygous deletion
methylated
point mutation
splice site mutation
other abnormal
fusion candidate
exon 2 deletion
b
HPV positive samples
n = 36
c
methylated samples
n = 47
d
heterozygous deleted samples
n = 91
e
samples with splice site mutations
mutation site
n = 2
median HPV for reference
f
homozygous deleted samples
n = 77
g
abnormal transcripts not otherwise specified
exon 2 deletion, n = 4
fusion candidate, n = 1
median HPV for reference
h
remaining samples
n = 24
i
medians for 3 major groups
hpv positive
methylated
homozygous deleted
normalized read depth ($\log_{10}$ plotting)
5000
1000
500
100
50
10
5
0
exon 1β
exon 1α
exon 2
exon 3

Figure S3.5. Alterations of *CDKN2A* gene structure, copy number, and expression of its protein coding transcripts p16INK4A and p14ARF. (a) Matrix documenting alterations of the *CDKN2A* locus as a function of HPV status including deletions (homozygous and heterozygous), methylation of the E1 alpha exon promoter, and single nucleotide substitutions (mutations). Splice site mutations are highlighted as samples with "abnormal transcript construction" as defined by SigFuge or MapSplice. HPV positive samples in general have no alterations in *CDKN2A* with the exception of a small number of predicted heterozygous deletions. However, we note that heterozygous loss as predicted by SNP chip is challenging, and there is no associated decrease in expression of *CDKN2A* in HPV positive samples predicted to have heterozygous deletion (red curves from panel (b)). Each of panels (b) through (i) represents a schematic of the *CDKN2A* gene locus which encodes 2 protein coding transcripts p16INK4A and p14ARF. The two proteins differ by their starting exons, with p16INK4A coded by exons 1α, 2 and 3, and p14ARF coded by exons 1β, 2 and 3. The exon structure is shown at the bottom of each panel. Plotted in each panel is normalized per-base expression along the gene for specific subsets of the 279 HNSCC samples. In each panel, subsets named by the panel title are shown as individual curves. The color of the curve relates to the colored row in panel (a) for that sample. In addition to the individual curves, each panel contains a bolded curve corresponding to the expression level of the median of a subset of samples. Panel (c) shows a striking pattern of exon E1 alpha loss in all samples classified as methylated (see Section S8) with overall retained expression of the exons of p14ARF. A small number of samples are indicated as having a coordinated single copy loss of *CDKN2A* (dark blue curve) or mutation (purple), but these too appear to be dominated by the expression pattern of E1 alpha silencing with retained p14ARF expression. Looking at the large set of heterozygously deleted cases (panel (d)) shows that the overall expression of the *CDKN2A* locus is lower than in the HPV positive samples as expected, and that the E1 alpha locus has median expression close to zero. In conclusion, these samples have less p14ARF and essentially no p16INK4A transcript expression. Panel (e) confirms the appearance of an abnormal transcript structure of p16INK4A associated with a splice site mutation. Panel (f) documents nearly absent expression along the entire *CDKN2A* locus, with the exception of a small number of clearly expressed samples which are likely false calls of deletion from the SNP chip data. Nonetheless these samples are appreciated as likely p16INK4A loss through either methylation and absent E1 alpha expression or mutation as indicated by purple and green traces. Panel (g) illustrates 5 cases in which abnormal transcripts were detected by either MapSplice fusion prediction or the SigFuge algorithm which would clearly produce a non-functional p16INK4A transcript. Panel (h) documents the remaining 24 cases in which no distinct *CDKN2A* alteration was identified from the genomic platforms. Strikingly, however, these samples appear to generally express *CDKN2A* transcripts that appear more consistent with genes with copy loss (such as in panel (d)) than intact genes (panel (b)). This suggests that even after identifying 90% (219 of 243) of HPV negative cases (n = 63 point mutations, n = 73 homozygous deletions, n = 33 heterozygous deletions, n=45 DNA methylation cases, n = 5 abnormally spliced transcripts) as "lost" there may still be additional cases of loss.

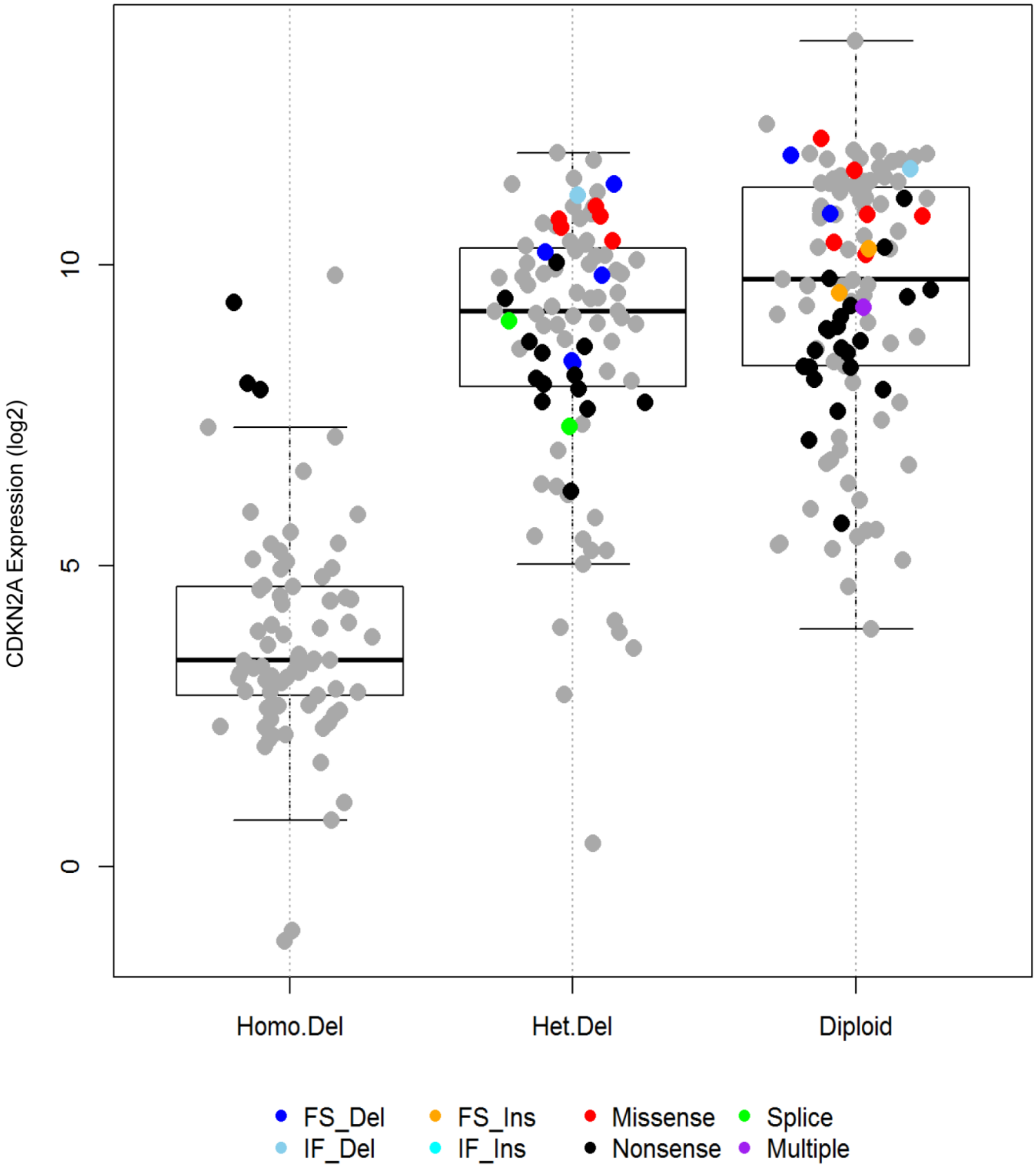


Figure S3.6. Integration of DNA mutation type, copy number, and gene expression for *CDKN2A*. Mutations are indicated as coded in the legend as follows: navy blue = frame shift deletion, orange = frame shift insertion, red = missense, green = splice site, sky blue = in frame deletion, cyan = in frame insertion, black = nonsense , and purple = multiple mutations in single sample. Gray color indicates wild type sequence. Overall the pattern is similar to that recently published for *CDKN2A* in LUSC. A clear association between homozygous deletion (Homo. Del) and decreased gene expression is observed, as expected for a driver tumor suppressor gene. In LUSC, samples with heterozygous deletions exhibited lower *CDKN2A* expression, as did

samples with nonsense mutations which are likely associated with nonsense mediated decay [31]. In contrast, higher *CDKN2A* expression levels were observed in LUSC samples with missense mutations. Although the effect size is not as large, similar trends are observed in HNSCC. Figure S4.2 part 1 shows the predicted coding impact of the *CDKN2A* mutations by transcript base position and functional domain.

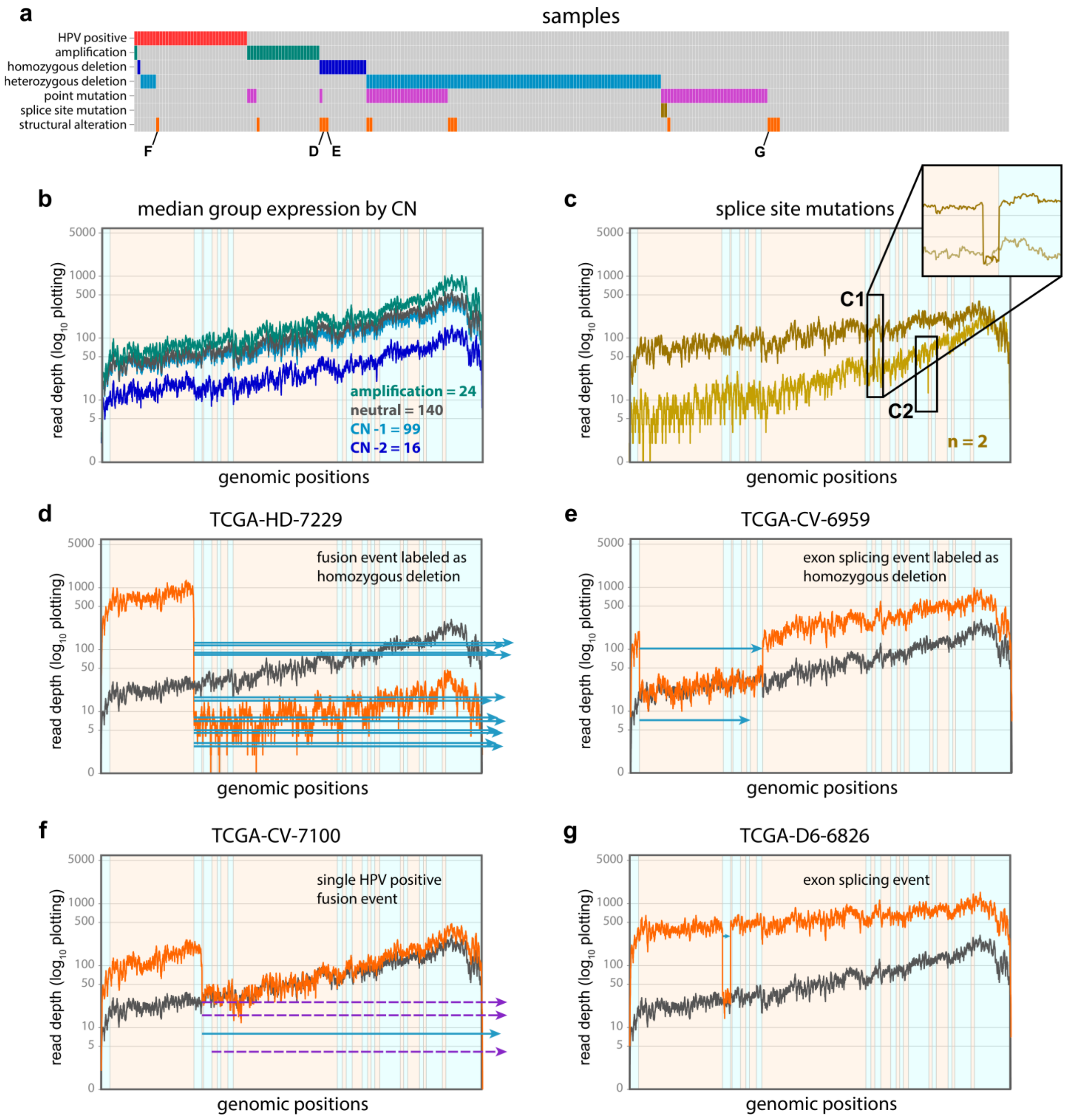


Figure S3.7. Alterations of *FAT1* gene structure, copy number, and expression. (a) Matrix documenting alterations of the *FAT1* locus as a function of HPV status for 279 HNSCC patients. Alterations shown include deletions (homozygous and heterozygous), single nucleotide substitutions (mutations), and structural alterations. Splice site mutations are highlighted as are samples with "abnormal transcript construction" as defined by SigFuge or MapSplice. HPV positive samples in general have no mutations and few alterations overall in *FAT1* with the exception of a small number of predicted heterozygous deletions. Panels (b) through (g) plot the per-base expression along the exons for several samples with notable mutations or structural alterations. In panels (b)-(g), alternating exons are colored blue and orange along the horizontal axes. (b) Consistent with a tumor suppressor phenotype, average gene expression is strongly associated with copy number. (c) Splice site mutations are documented to have the expected impact on transcript composition. The locations of the mutations along the transcript are visible as dips in expression near the splice junction (C1, C2). A zoomed-in view of the region surrounding the splice site mutation is shown for one sample (C1). Even more striking than for *CDKN2A*, *FAT1* demonstrates a large variety of structural abnormalities, in particular challenging the definitions of deletions, splices, and fusions. Panel (d) is a case where copy number analysis by SNP chip identifies the sample as homozygously deleted, yet gene expression analysis documents high expression by conventional measures (which averages counts of mapped reads across the locus). Figure S3.8 indicates this sample as one of 2 open red circles (labeled with *) in the homozygous deletion column, clearly demonstrating *FAT1* gene expression above the median for the cohort. The paradox is resolved when RNA coverage is displayed across the gene, and fusion detection algorithms are applied revealing an intra-chromosomal fusion (both donor and acceptor components on chromosome 14). Panel (e) documents a similar event with the exception that the deletion event is much smaller, so small in fact that it occurs entirely within the gene. Accordingly, the RNA event is identified not as an intra-chromosomal fusion, but as an alternative splice from the MapSplice splicing file rather than the predicted fusions. Panel (f) is the only *FAT1* fusion involving an HPV positive sample and one of only 2 inter-chromosomal fusions for the gene, interestingly both with acceptor sites on chromosome 15. The 2 acceptor sites were intragenic although separated by less than a megabase on chromosome 15. Panel (g) represents an exon skipping event, again captured in the MapSplice splicing file rather than the fusion file. There is no splice site mutation detected in the sample. Whether there is a small intragenic deletion not appreciated by SNP chip or other mechanism causing the splice is unknown. This type of exon skip without a mutation or detected deletion is seen in *CDKN2A* (Figure S3.5 panel (g)) and labeled as "exon 2 deletion," suggesting that these types of events may be common.

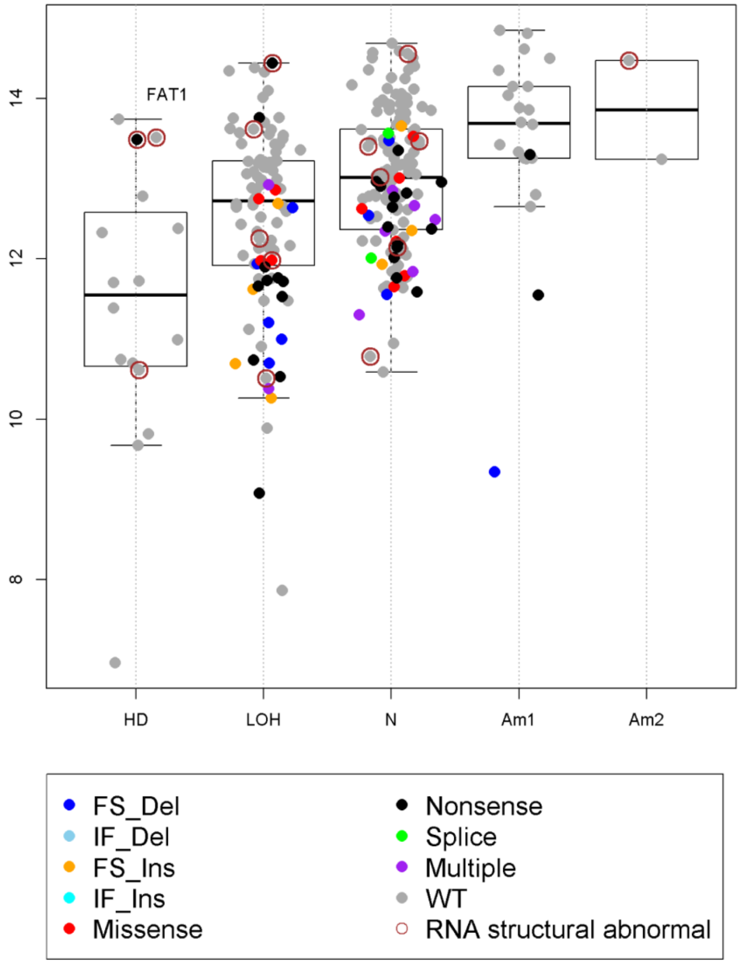


Figure S3.8. Integration of DNA mutation type, copy number and gene expression for *FAT1*. Discrete DNA copy number values for *FAT1* are shown on the x-axis, while gene expression measurements are shown on the y-axis. Mutations are indicated as coded in the legend as follows: navy blue = frame shift deletion, sky blue = in frame deletion, orange = frame shift insertion, cyan = in frame insertion, red = missense, black = nonsense , green = splice site, , and purple = multiple mutations in single sample. Gray color indicates wild type sequence. A clear association between single copy DNA loss (aka heterozygous deletion, written LOH), 2 copy loss (aka homozygous deletion, written HD) and gene expression is observed, as expected for a driver tumor suppressor gene. Moreover, few samples exhibit low copy gains (defined as 1 by the GISTIC algorithm) or high copy gains (defined as 2 by the GISTIC algorithm), written Am1 and Am2, respectively. An open circle represents evidence of predicted structural alteration from RNA or low pass whole genome sequencing as in Figure S3.7. Many samples with a single copy loss of *FAT1* show evidence of a second hit through sequence alteration or structural alteration. Most of the cases with multiple mutations occur in copy neutral samples. The pattern that missense mutations are expressed at higher levels than nonsense mutations, explained by the phenomenon of nonsense-mediated decay is seen as in Figure S3.6 although less prominently.

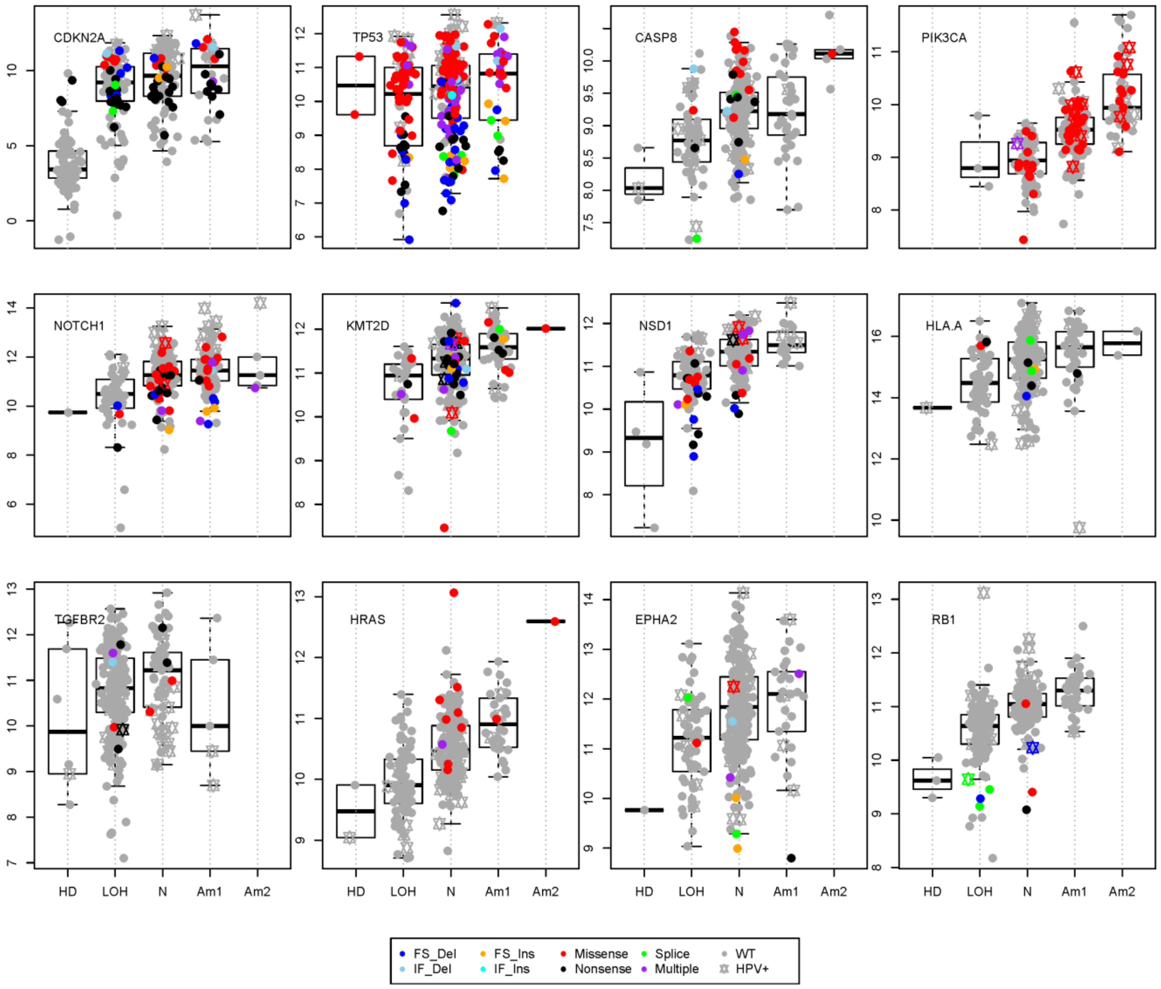


Figure S3.9, part 1. Integration of DNA mutation type, copy number, and gene expression for predicted driver genes relevant to HNSCC. Discrete DNA copy number values for select genes are shown on the x-axis, while gene expression measurements are shown on the y-axis. Mutations are indicated as coded in the legend as follows: navy blue = frame shift deletion, sky blue = in frame deletion, orange = frame shift insertion, cyan = in frame insertion, red = missense, black = nonsense , green = splice site, and purple = multiple mutations in single sample. Gray color indicates wild type sequence. A star indicates HPV(+). DNA copy number status is coded as follows: HD = homozygous deletion, LOH = heterozygous deletion, N = copy neutral, Am1 = single copy gain, Am2 = multiple copy gain.

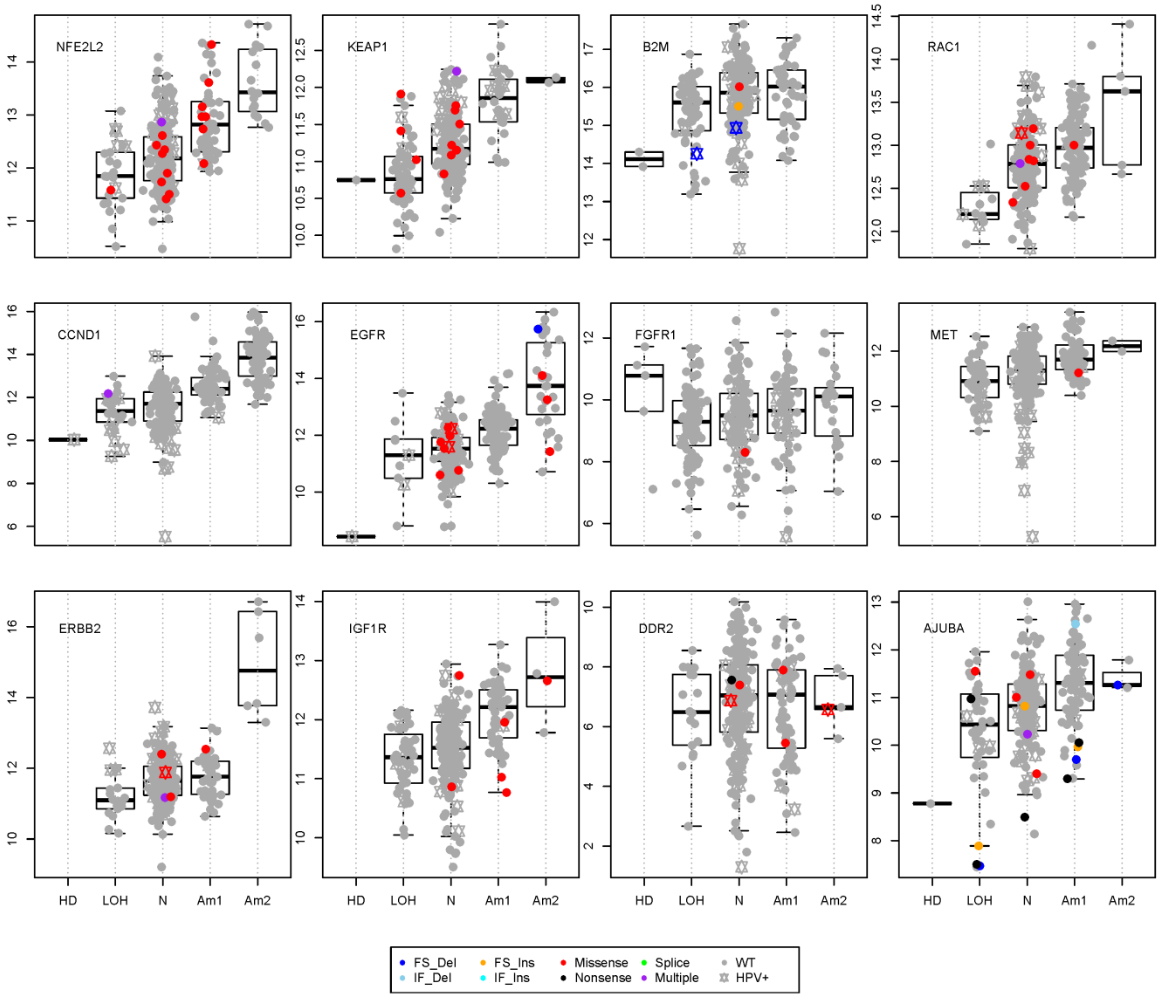


Figure S3.9, part 2. Integration of DNA mutation type, copy number, and gene expression for predicted driver genes relevant to HNSCC

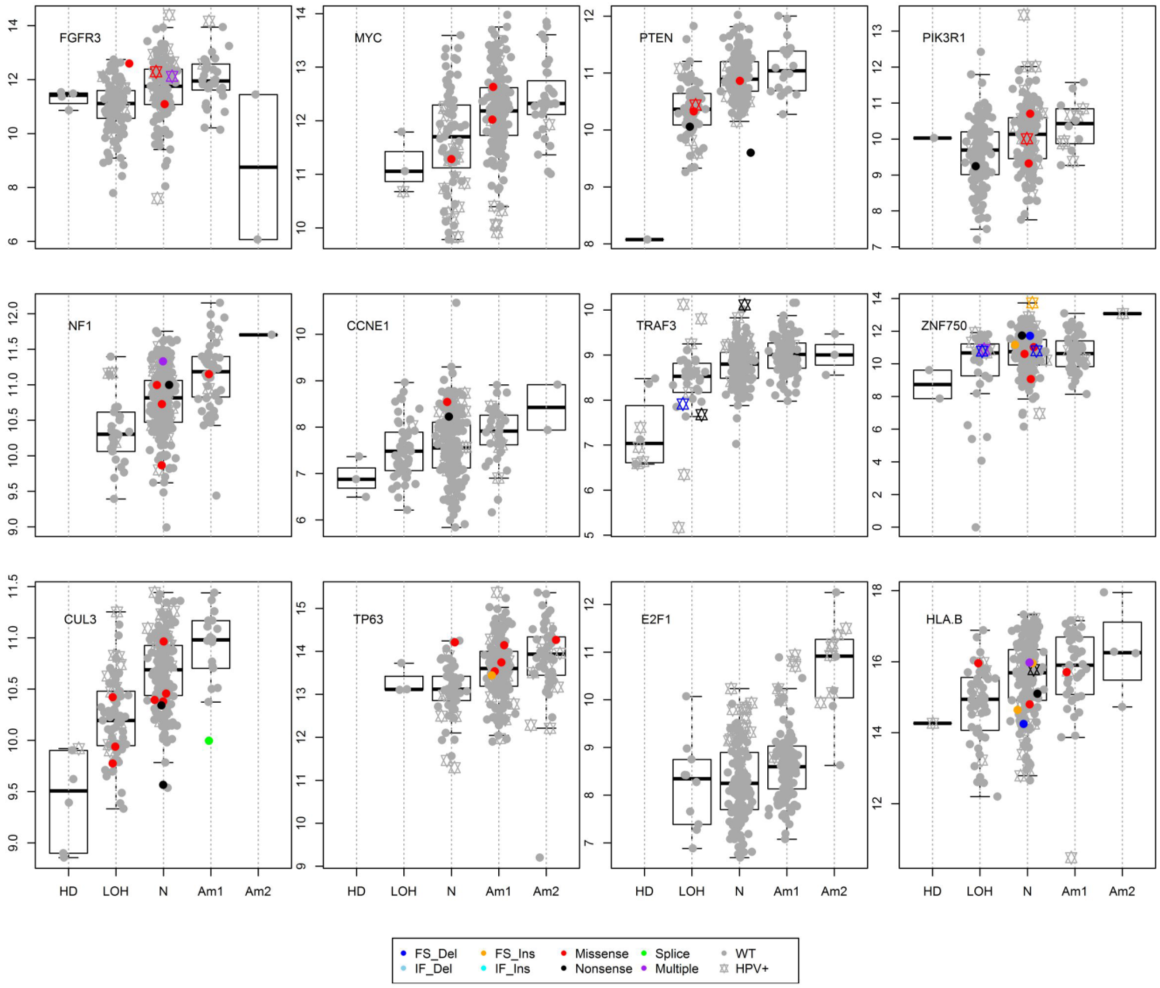


Figure S3.9, part 3. Integration of DNA mutation type, copy number, and gene expression for predicted driver genes relevant to HNSCC.

Figure S3.9. Integration of DNA mutation type, copy number, and gene expression for predicted driver genes relevant to HNSCC. In tumor suppressor genes a positive correlation between copy number and gene expression is observed, with greater tendency to be deleted than amplified. Many samples with a single copy loss of the gene show evidence of a second hit through mutation. Most of the cases with multiple mutations occur in copy neutral samples. The pattern that missense mutations are expressed at higher levels than nonsense mutations, explained by the phenomenon of nonsense-mediated decay is common. Oncogenes demonstrate a different pattern with few nonsense mutations, greater tendency to be amplified with few deletions, and again a positive correlation between gene expression and copy number.

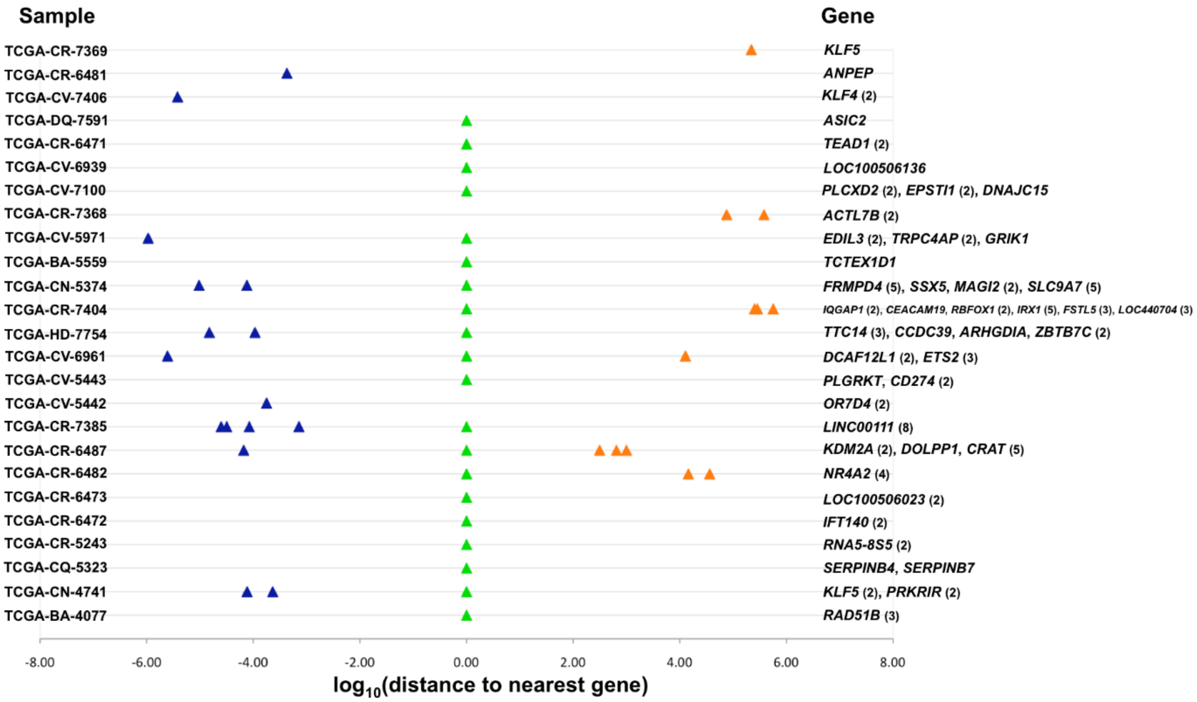


Figure S3.10. Distribution of HPV integration breakpoints across the host genome. 102 viral-host genome integration breakpoints per each of 25 HPV+/Integration+ tumors plotted versus the $\log_{10}$(distance to the nearest human gene). Zero distance denotes integration event within a gene, and these are colored green. Negative and positive distance values mark integrations happened upstream and downstream, and these are colored blue and orange, respectively. 20 (80%) tumors have at least one integration breakpoint within a gene. Number of breakpoints related to the certain gene is specified in parentheses. In one case, TCGA-CN-4741, integration occurs 13 kb upstream of *KLF5*. However, in the second case, TCGA-CR-7369, integration occurs 219 kb downstream of *KLF5*. Although *KLF5* is the nearest gene in both cases, the distance between the two integration sites is too large to classify them as recurrent.

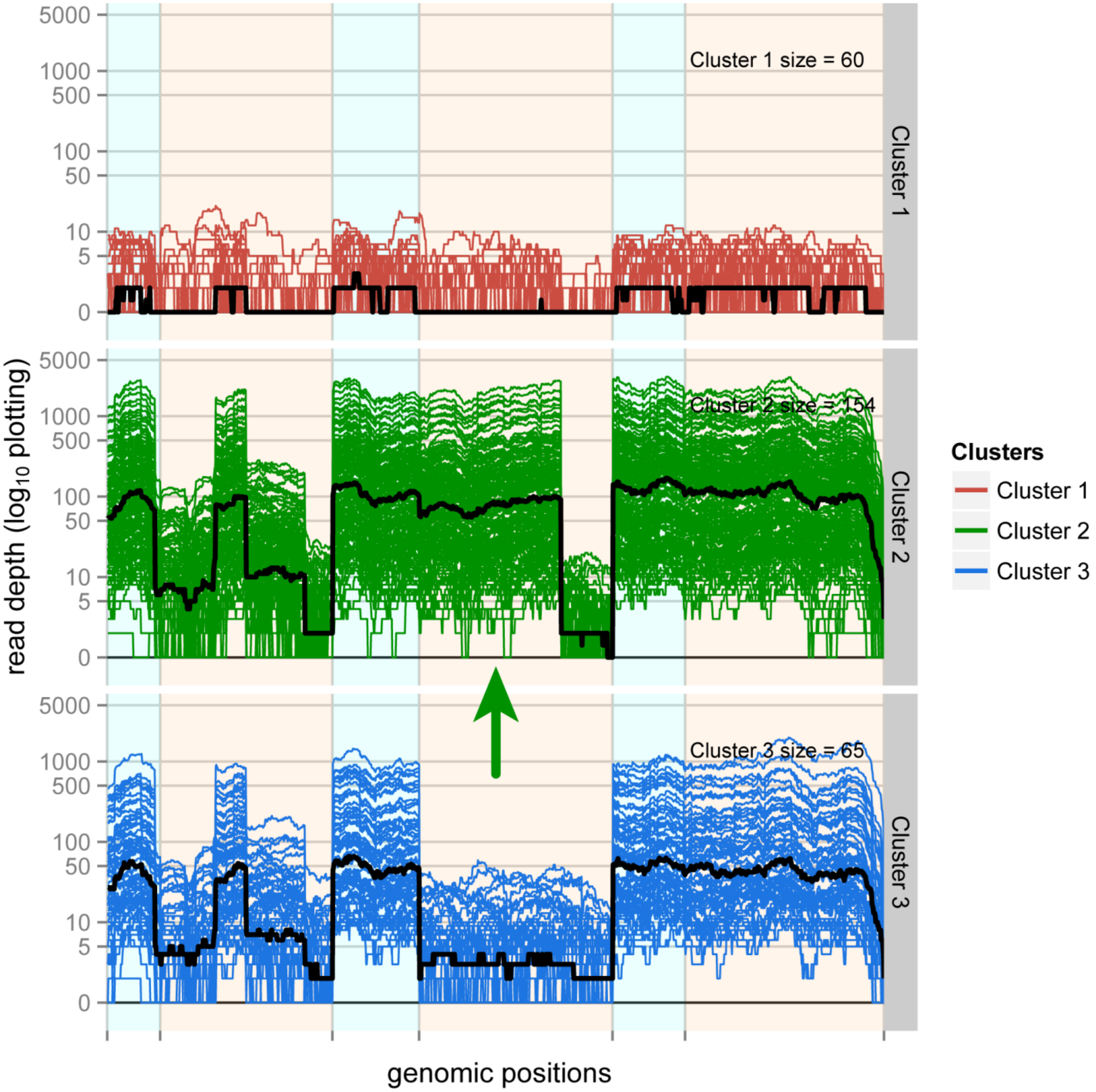


Figure S3.11. The *KLK12* gene documents recurrent alternate transcription in HNSCC. The *KLK12* locus is shown as a function of transcript position from 5' to 3' and as a function of cluster identified by the SigFuge clustering algorithm. Exons are shown as alternating blue and red shaded positions. Each sample is a curve on the figure. The height of the curve represents normalized RNA coverage at that position. Regions of low coverage in all samples represent predicted gene territory for which measured RNA registered no coverage. Three clusters are identified: a red cluster (n=60) with essentially no expression of the gene, a green cluster (n=154), and a blue cluster (n=65). The difference between green and blue clusters is an absent exon 4 in the blue samples as indicated by the green arrow.

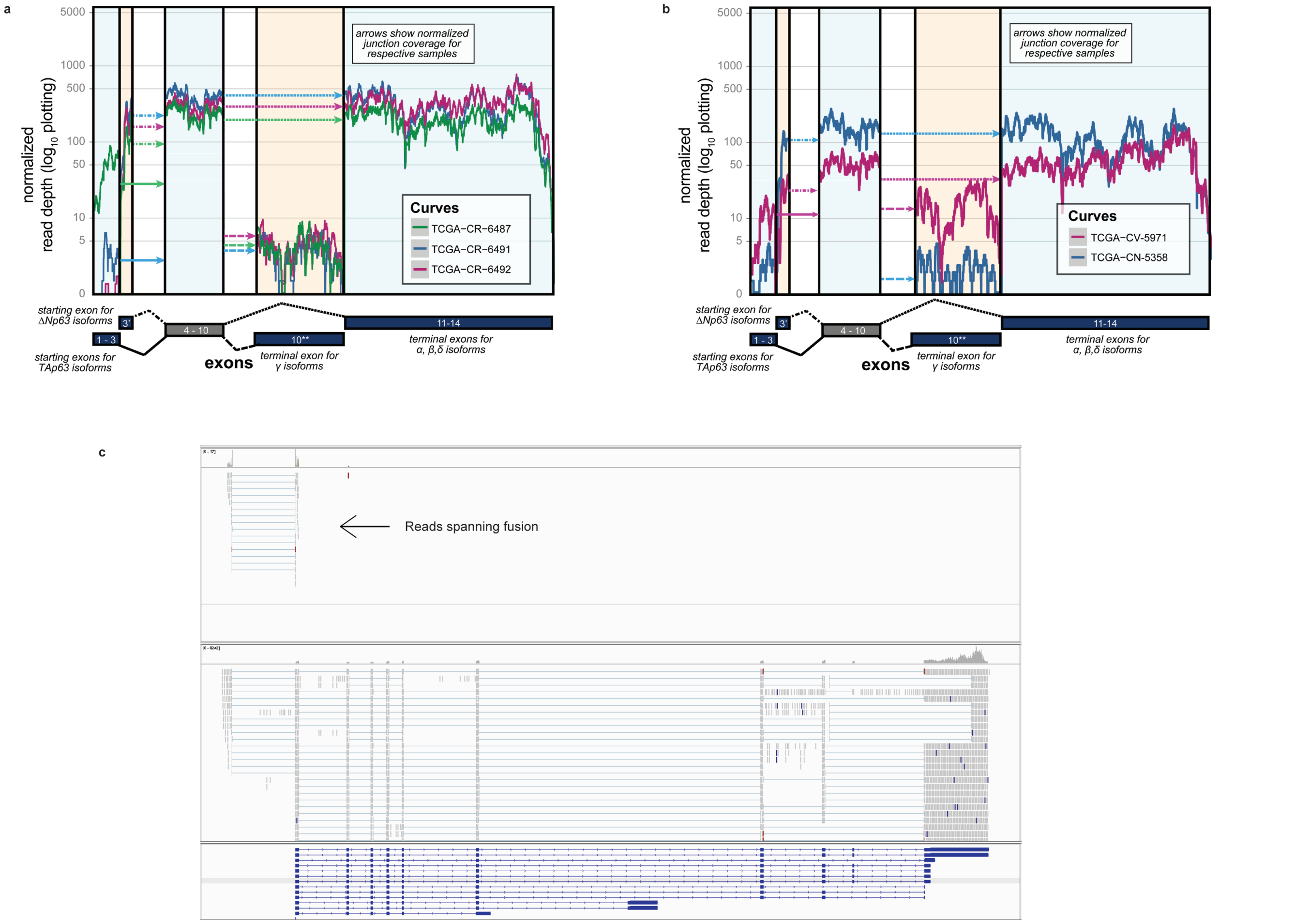


Figure S3.12. Heterogeneous *TP63* isoform usage in HNSCC. (a) The gene *TP63* is described by 2 primary isoforms depending on transcription start site, either TAp63, starting at exon 1 or DeltaNp63 starting at exon 3 in this representation of the gene (n.b. alternate gene annotations of TP63 exist that sometimes annotate the DeltaNp63-specific exon as “exon 5”). Both the TAp63 and DeltaNp63 are further categorized into at least 5 additional isoforms depending on 3’ exon splicing. These different isoforms are shown in simplified gene models given at the bottom of the figure. A dash-dotted versus solid line indicates exon junctions with more than one donor or acceptor site. Arbitrary space is introduced after exon 4 and exon 11 for easier visualization of the splicing. The normalized per-base expression is plotted for two samples along the exons of *TP63*, representing the variation of *TP63* expression among the HNSCC samples. Shown in panel (a) are examples with high expression of TA (green only) and DeltaN (green, blue, and magenta) isoforms. Panel (b) documents a case (magenta) in which the TA and DeltaN isoforms are expressed at nearly equal amounts contrasted to the typical pattern represented by the blue example with low TA expression and high DeltaN. Unlike samples from panel (a) and the magenta sample from panel (b), however, the blue sample shows essentially no expression of the gamma isoform. Panel (c) shows a case (TCGA-BA-5152) in which there is evidence for an intra-chromosome fusion of chromosome 3q (chr3 position189315027 fused to exon 5 of *TP63*). The top portion of (c) documents bridging reads and the bottom documents all reads mapped to the predicted fusion. While the resulting transcript has not been characterized functionally, it would include the DNA binding domain of the DeltaN isoform. TCGA-BA-5152 is notable for two additional characteristics. It is in the minority of cases (approximately 25%) in which there is no measured amplification of the *TP63* locus by the GISTIC copy number analysis algorithm, suggesting that fusion could be an alternate mechanism of

activation of the locus. TCGA-BA-5152 also contained a *TP63* mutation (E248K) of unknown significance. Of the 10 samples with fusions, mutations, or both, only 1 occurred in the context of a GISTIC copy number estimate of 2 (highly amplified) which supports the observation that somatic alterations might substitute for amplification in some cases. Four of eight *TP63* mutations from the 279 HNSCC cases have previously been observed in the COSMIC database lending some support to the alterations being functional driver events.

S4: DNA sequencing

S4.1. Exome sequencing, high-pass whole genome sequencing, and data processing

Whole exome sequencing (n=279) and high coverage whole-genome sequencing (30X, n=29) were performed as previously described[31]. Briefly, 0.5-3 micrograms of DNA from each sample was used for library preparation, which included shearing and ligation of sequencing adaptors. Exome capture was performed using the Agilent SureSelect Human All Exon 5Mb kit. Captured DNA was sequenced using the Illumina HiSeq platform, and paired-end sequencing reads were generated for each tumor and matched normal sample. Initial alignment and quality control were performed using the Picard and Firehose pipelines at the Broad Institute.

Picard generates a single BAM file for each sample that includes reads, calibrated quantities, and alignments to the genome. Firehose represents a set of tools for analyzing sequencing data from tumor and matched normal DNA. The pipeline uses GenePattern16 [43] as its execution engine, and performs quality control, local realignment, mutation calling, small insertion and deletion identification, rearrangement detection, and coverage calculations, among other analyses. Complete details of this pipeline can be found in Stransky et al. [9] or www.broadinstitue.org/cancer/cga. The results were summarized according to the mutation annotation format (MAF) specification, and the MAF contains this summary. Identification of inter-chromosomal and large intra-chromosomal structural rearrangements was performed using the dRanger algorithm [43]. Candidate rearrangements were identified as groups of paired-end reads which connected genomic regions with an unexpected orientation and/or distance. When possible, breakpoints were mapped to basepair resolution using BreakPointer [44]. In 279 samples, a total of 12,159 synonymous somatic variants, 37,061 non-synonymous variants, and 2,579 germline single nucleotide variants from dbSNP [45] were detected.

Mutation significance analysis

Mutation significance was performed using the MutSigCV algorithm[46]. In brief, this algorithm takes into account recurrence of mutations, nucleotide context, gene expression, replication time, and somatic background mutation rate. The overwhelming majority of genes tested by MutSig produce a q-value equal to 1, as evidenced by the fact that only 19 genes had MutSig q-values less than 1. Genes with a q-value less than 0.1 were deemed significant. To identify less common mutations with known roles in cancer that may have been missed by our computational approach, we extracted genes within the Cancer Gene Census[47] and utilized the Mutsig algorithm taking into account only genes within this previously annotated set of genes with known roles in cancer. We also performed Mutsig independently on all four classes of expression subtypes, on HPV(+) and HPV(-) tumors independently and on samples originating in the oral cavity, larynx and oropharynx to identify genes enriched for mutation in each of these groups. For example, inactivating mutations of *TGFBR2* were found predominantly in oral cavity tumors, consistent with its role in promoting squamous tumorigenesis in mouse models [48]. These results are summarized in Data File S4.1 Tab 1. We also include for genes found to be significantly mutated in a specific anatomic site or in HPV(+) or HPV(-) tumors p values and FDR-adjusted p values for enrichment by site or HPV status (Tabs 2 and 3). Using the methods of Lawrence and colleagues [46], which combine output from MutSigCV and two alternate versions of MutSig, genes found to be significantly mutated by site were identified, including *PTEN*, *CUL3*, and *RB1* (Data File S4.1 Tab 4). Mutations from selected genes, primarily those from the MutSig lists and other suspected driver genes in HNSCC, are displayed as a function of transcript position and functional domain (Figure S4.2)

S4.2 Mutation validation

Targeted resequencing of selected mutations for validation was performed by PCR using a microfluidic device (Fluidigm). PCR primers with Fluidigm-compatible tails were designed to flank sites of interest and produce amplicons of 200 bp +/-20bp. 20 – 50 ng of each DNA sample was mixed with oligonucleotides containing Illumina adapter sequences, a sample-specific molecular barcode and a sequence complementary to the primer tails. This mixture was used as the PCR template for each sample amplified on the Fluidigm access array. This method allowed the use of universal PCR primers while ensuring all amplicons for a given sample received the same sequencing barcode. PCR was performed on the Fluidigm access array according to manufacturers' instructions. Barcoded libraries were recovered for each sample in a single collection well on the Fluidigm access array, quantified using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA) and concentrations normalized across libraries. Libraries were loaded on the Illumina MiSEQ instrument and sequenced using paired end 150bp sequencing reads (Figure S4.1). Targeted re-sequencing of 394 unique regions across 278 tumor/normal pairs with sufficient DNA was performed to validate point mutations (Figure S4.1). The validation rate was 99% at the 94% of sites with adequate coverage (those with 95% power to detect a mutation).

S4.3. Low pass whole genome sequencing

Library construction

Low pass WGS (5X, n=98) was performed as previously described [49]. Between 500 and 700 ng of each gDNA sample were sheared using Covaris E220 to about 250 bp fragments, then converted to a paired-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturers' protocols. All libraries were sequenced on HiSeq2000 using one sample per lane, with the paired-end 2 x 51bp setup. Tumor and its matching normal were usually loaded on the same flowcell. Average sequence coverage was 6.49X, read quality was 38.5 and 94.6% of the reads were mapped. Raw data were converted to the FASTQ format and BWA alignment was used to generate BAM files.

Identification of copy number variants

To characterize somatic copy number alterations in the tumor genome, we applied BIC-seq [50], an algorithm we have developed previously. Briefly, we first counted the uniquely aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both the fit and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio difference smaller than 0.1 (log2 scale) between tumor and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

Discovery of rearrangements with BreakDancer and Meerkat

We used two programs to detect structural variation with increased sensitivity, BreakDancer [51] and Meerkat [52]. The first step in BreakDancer requires a configuration file of each BAM file for each tumor pair with the bam2cfg.pl perl module of the program. The perl module BreakDancerMax.pl is then run on the configuration file to call structural aberrations (SAs) in the tumor and normal files. The set of SA calls from each tumor sample is filtered by the calls from its matched normal to remove germline variants. Structural aberrations were also detected by Meerkat, which requires at least two discordant read pairs supporting each event and at least one read covering the breakpoint junction. To be more conservative, variants detected from tumor genomes were filtered by the variants from all normal genomes to remove germline events and were also filtered out if both breakpoints fell into simple repeats or satellite repeats. The final call has to fulfill the following: (i), the read identified has to span the breakpoint junction and hit the predicted breakpoint region uniquely by BLAT, or (ii), the mate of the read spanning the breakpoint junction is mapped near the predicted breakpoint. For our analysis, we have used the union of calls from the two algorithms (Data File S4.2). On average, 24 structural aberrations per tumor were detected from low pass WGS (n=98).

Validation of rearrangement hits

We required two or more discordant pairs to call an SA in order to reduce the number of false positive calls. In Meerkat, we also sought to find a read that spans the breakpoint. Such a read not only confirms our assertion of an SA but also provides a precise location at which the break occurs; however, lack of this split read does not necessarily indicate a false positive, as the sequencing coverage is low in our data. In a subset of cases where we could not find a read that contains the SA junction, we attempted to PCR amplify the junction fragment and subject it to Sanger sequencing. Finally, we also used RNA-seq data where available to assess if novel junctions between two genes can be found. Based on all three methodologies, we estimate that the accuracy of SA calling is accurate to a minimum level of 60%.

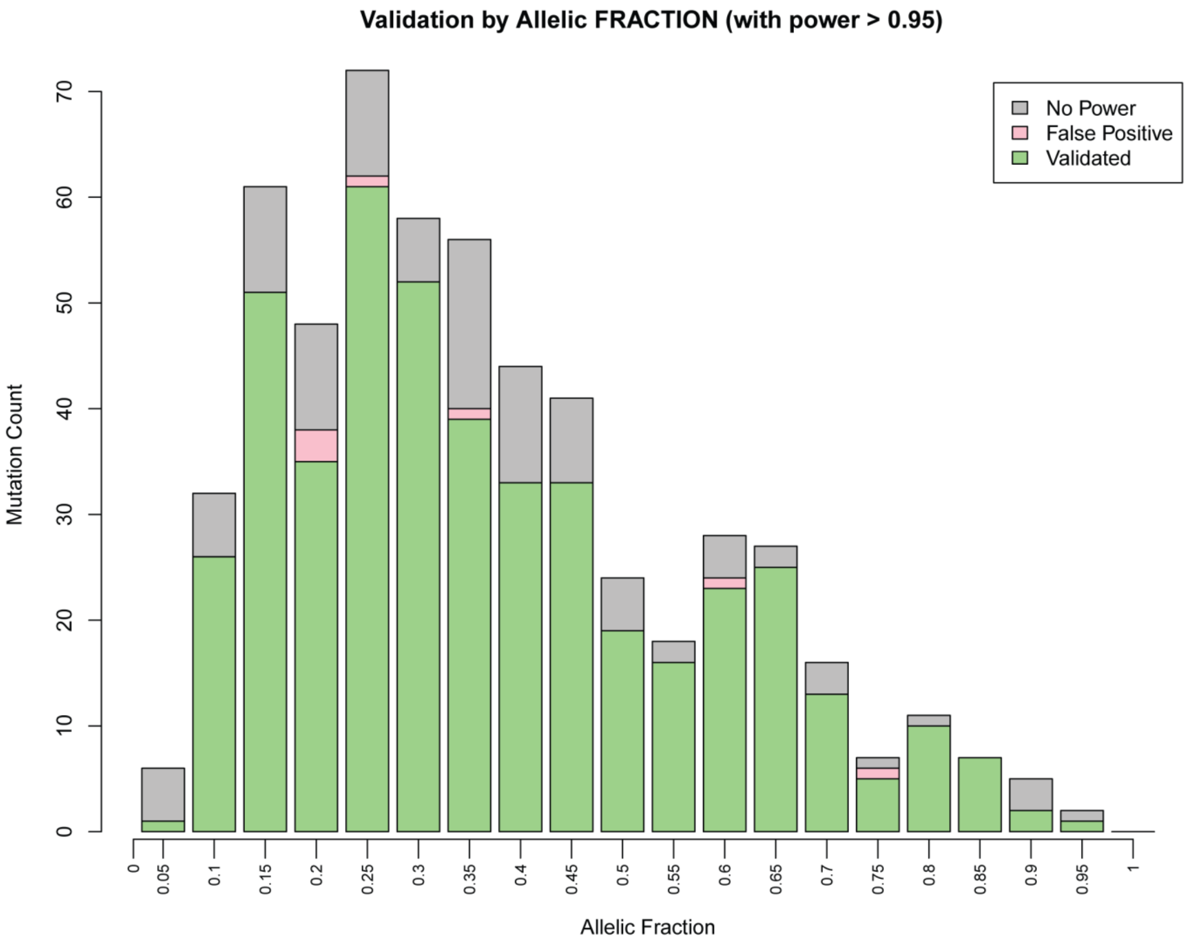


Figure S4.1. Mutation validation counts by allelic fraction for HNSCC. Validation of single nucleotide variant (SNV) calls in WES by PCR-based target enrichment. Shown are the number of mutations in which validation of mutations was attempted and the allelic fraction of the mutations in the original sequencing to the nearest 5%. Validated mutations are shown in green and mutations which were not validated in pink. Sites in which inadequate coverage of the target to validate or invalidate the event with 95% power are depicted in gray.

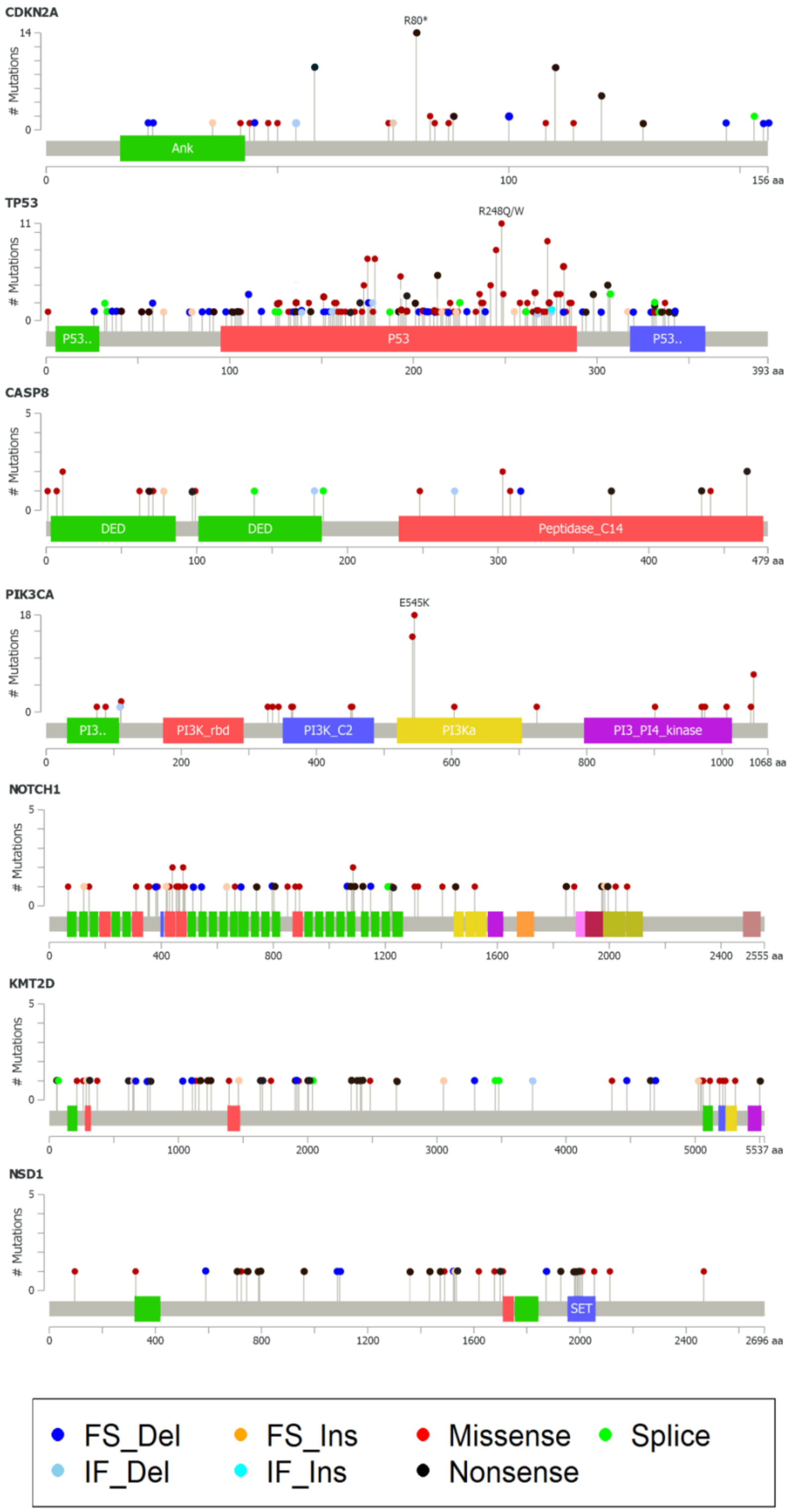


Figure S4.2, part I. Predicted coding impact by transcript base position and functional domain for selected genes.

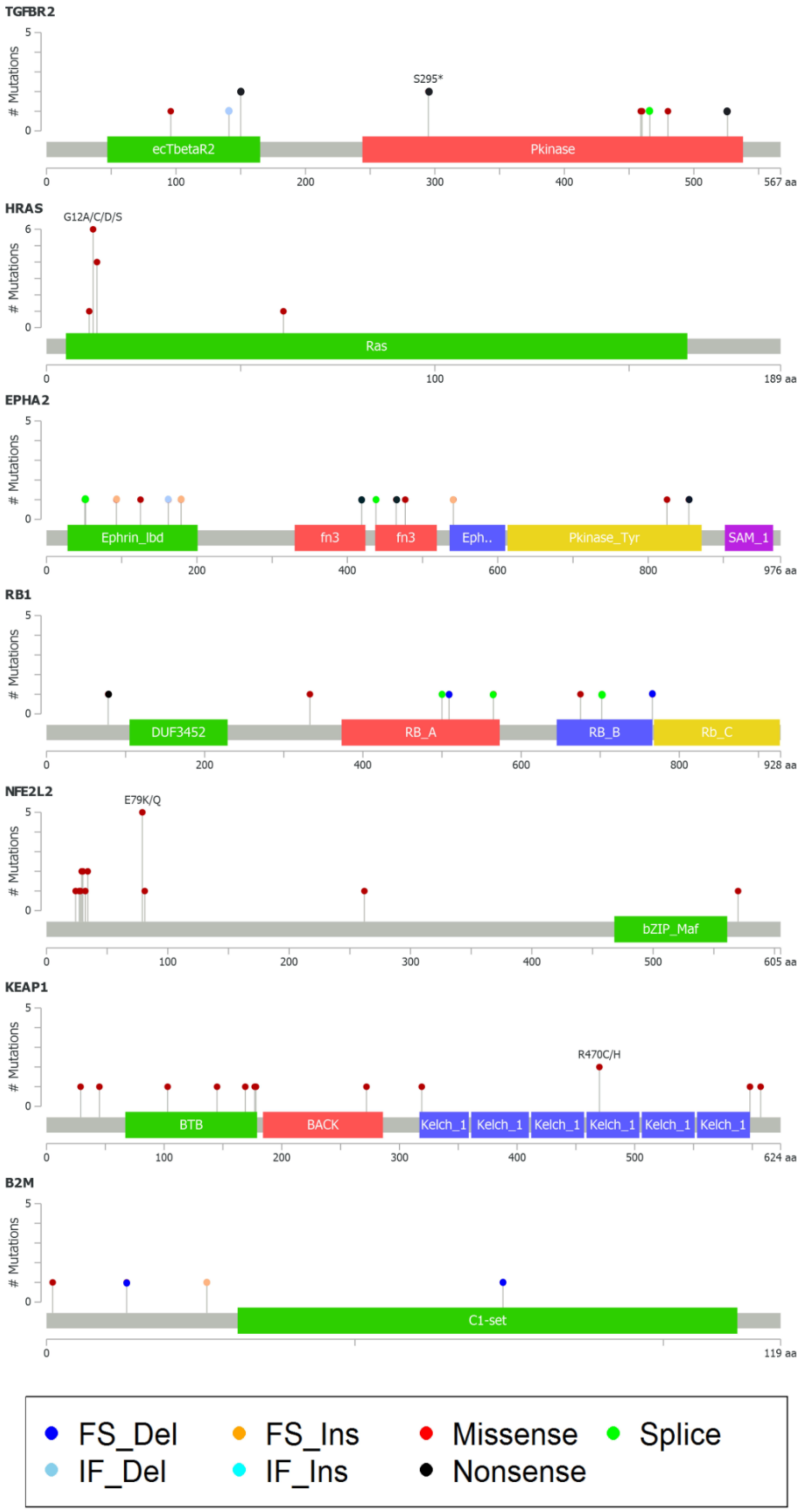


Figure S4.2, part 2. Predicted coding impact by transcript base position and functional domain for selected genes.

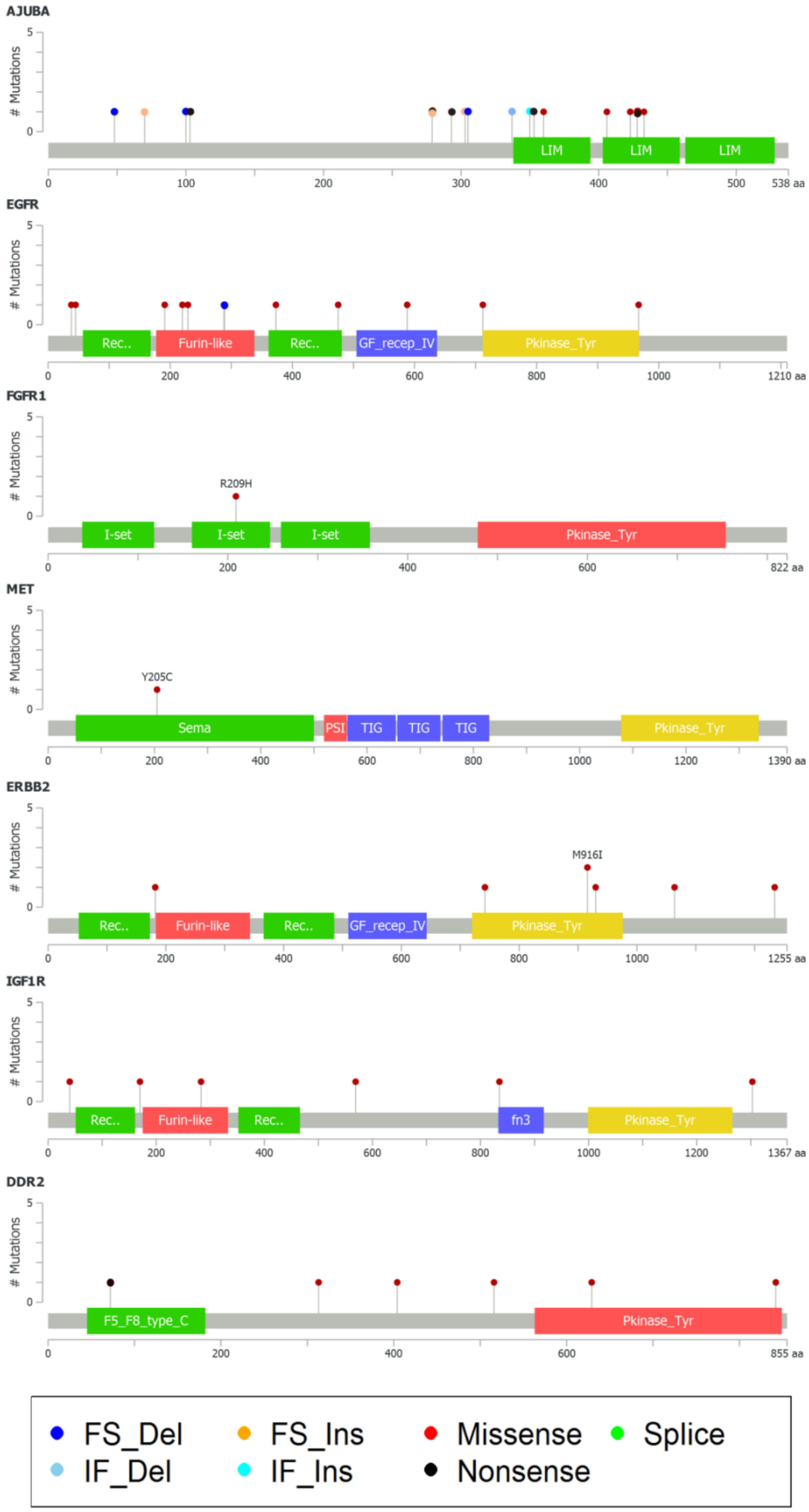


Figure S4.2, part 3. Predicted coding impact by transcript base position and functional domain for selected genes.

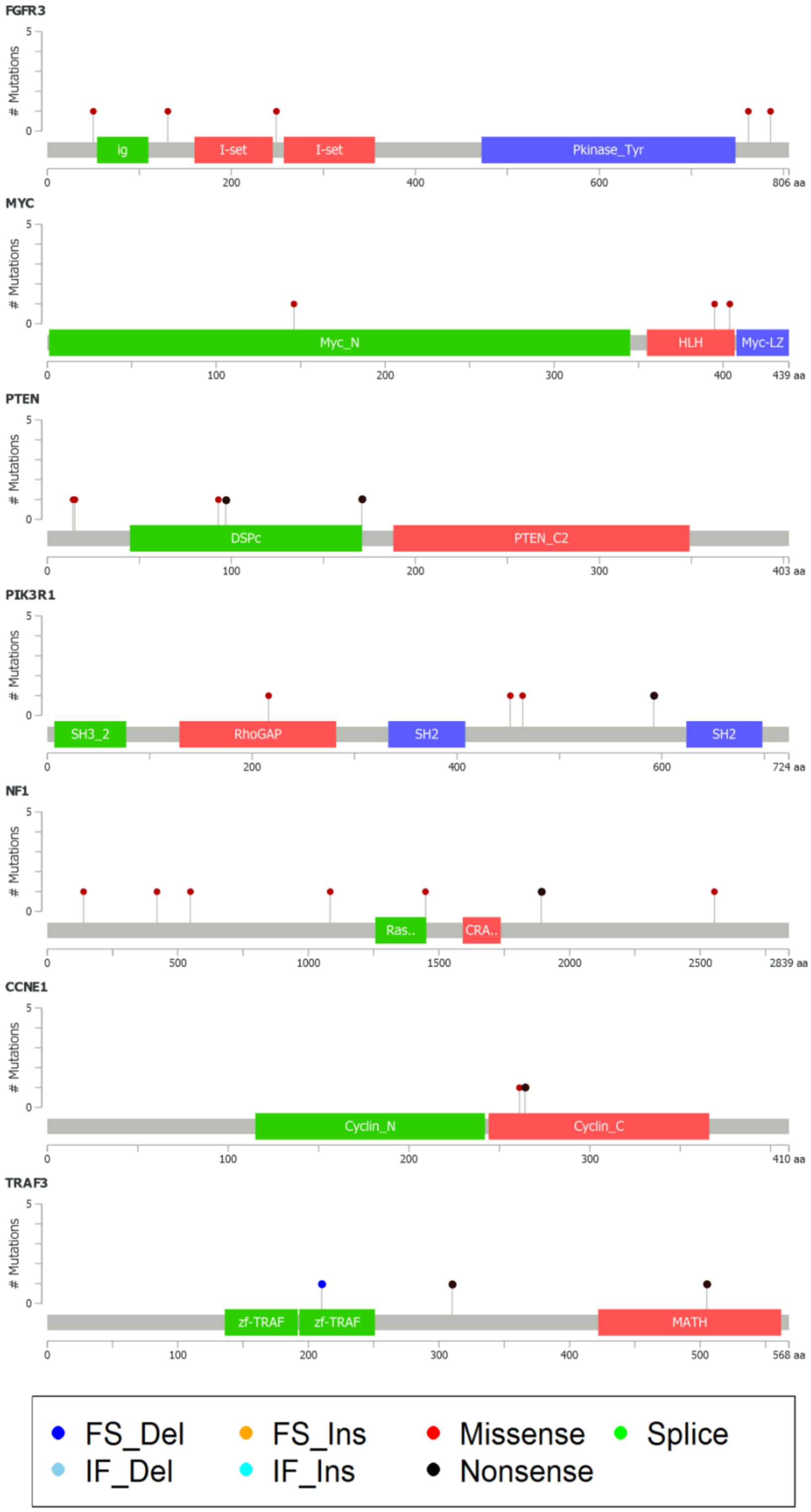


Figure S4.2, part 4. Predicted coding impact by transcript base position and functional domain for selected genes.

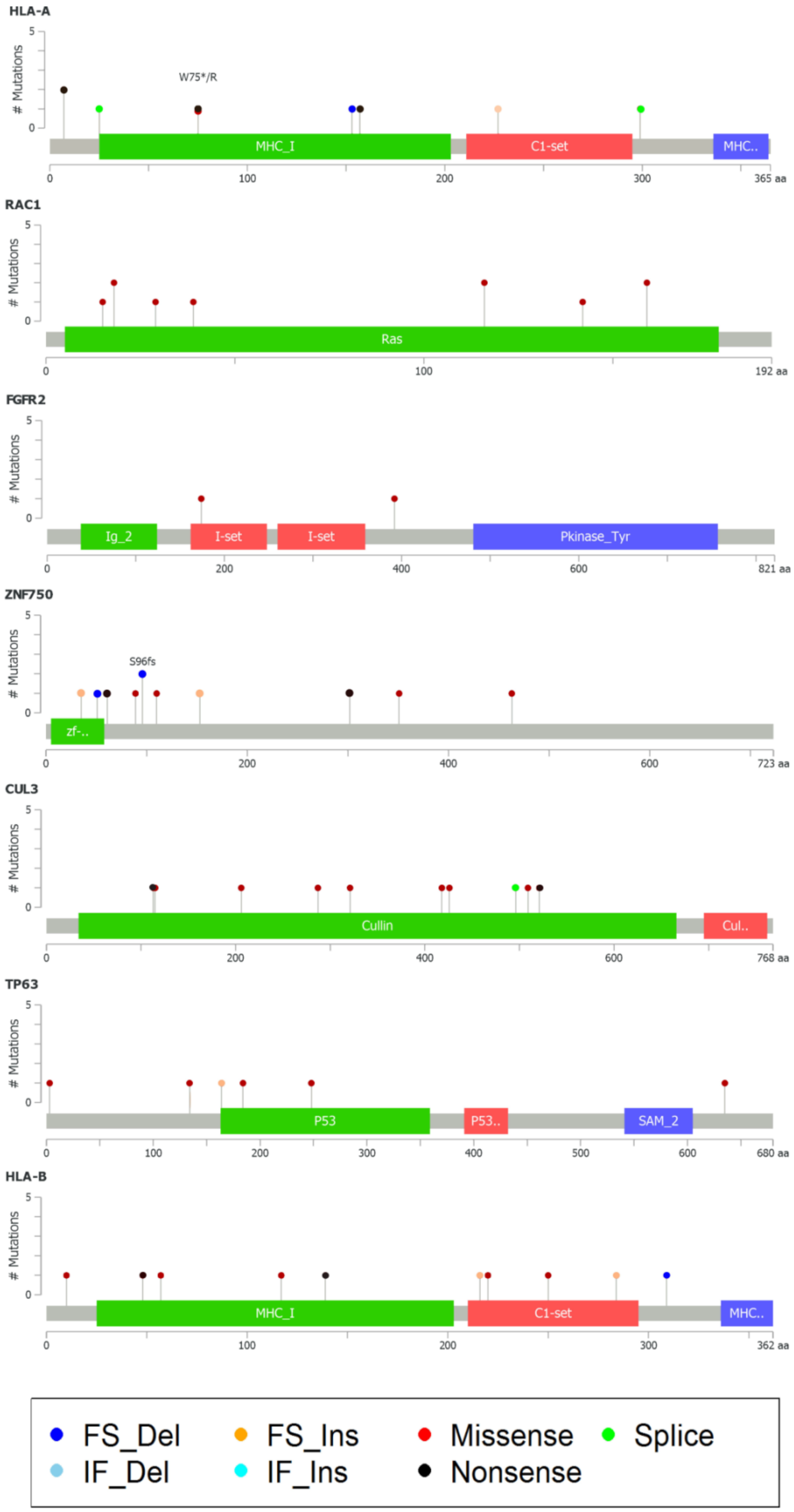


Figure S4.2, part 5. Predicted coding impact by transcript base position and functional domain for selected genes.

Figure S4.2. Predicted coding impact by transcript base position and functional domain for selected genes. Gene diagrams are shown schematically based on RefSeq [53] canonical isoforms and functional domains annotated in PFAM 27.0 [54]. The predicted impact of mutations detected from whole exon sequencing is

shown by a colored circle and stick (a.k.a. lollipop). Mutations from the MAF [37] are shown. The color code is as follows: navy blue = frame shift deletion, sky blue = in frame deletion, orange = frame shift insertion, cyan = in frame insertion, red = missense, black = nonsense, and green = splice site. The height of the lollipop is proportional to the number of mutations at a position. Lollipop plots add information beyond statistical significance by displaying the nature and distribution of alterations across the gene. For tumor suppressor genes such as *NSD1*, the proportion of nonsense mutations is relatively high across the gene, indicating many events producing truncated and likely non-functional proteins. *NSD1* loss has been associated with sporadic non-melanoma skin cancers [55], echoing the dual roles of *NOTCH1* as an oncogene in leukemia and a tumor suppressor in HNSCC and other cancers. Mutations occurring in multiple patients in the same position in a gene are suggestive of key regulatory positions in the transcript which can be activating and oncogenic, such as in *PIK3CA* or loss of function such as in *CDKN2A*. The figures shown above were created using cBioPortal [56,57] with manual coloring of the mutations. Interested users are referred to cBioPortal for additional details.

Of note, while *EGFR* demonstrates mutations in approximately 5% of samples, these are generally not in the kinase domain such as seen in lung adenocarcinoma and none were recurrent within the dataset. However, review of the COSMIC database reveals at least 2 cases where identical somatic substitutions have previously been reported including in diseases such as glioblastoma: A289T in sample TCGA-CV-7252 and F712L in TCGA-CV-7427. In other cases mutations in identical transcript positions have been reported, although with a different residue such as I475V in TCGA-CN-4742. The recurrent nature of these mutations, in particular A289T lends support to the biologic relevance of the alterations. Although no link between *EGFR* alterations and the efficacy of EGFR inhibitors has been established, it is interesting to note that the 16% frequency of *EGFR* mutations and amplifications in HPV(-) tumors was similar to the reported efficacy of EGFR inhibitors [58].

S5: Molecular subtypes and subset analyses

S5.1 Detection of previously validated gene expression subtypes in HNSCC and correlation with lung squamous cell carcinoma

Previously validated gene expression subtypes of HNSCC, termed basal, mesenchymal, atypical, and classical [59,60], were detected in the TCGA HNSCC cohort. RSEM values were log2 transformed, genes with missing values were removed, and the resulting expression measurements were median centered by gene. Using previously published predictor centroids and methods [60], each TCGA tumor specimen was assigned an expression subtype using a nearest centroid predictor. Not all genes from the prior predictor were available thus the predictor was adjusted from 838 genes to the overlap set of 728 (Figure S5.1, top middle panel). Using this overlap gene set, a comparison cohort was evaluated which had previously been assigned class labels with the full 838 predictor genes (Figure S5.1, top left panel). The expression patterns in the subtypes were highly concordant across the two cohorts, indicating that the subtypes provided similar stratifications of the subjects. Detection of gene-sample clusters in independent sample sets has been used as a component of the empiric definition of "cluster validation"[61]. Prior work has also suggested a correlation of HNSCC gene expression subtypes to TCGA lung squamous cell carcinoma (LUSC) expression subtypes, and using similar methods we assessed the correlation using TCGA HNSCC subtype data, again confirming similar patterns (Figure S5.1, top right panel). To further empirically assess the quality of these subtype detections similar to earlier studies, expression of selected predictor genes (Figure S5.2) were compared between the TCGA and a previously published cohort[60]. Additionally, the overall similarity between HNSCC and LUSC was further evaluated using a previously described technique[60]. The 728 predictor genes were used to define centroids for each of the 4 classes of HNSCC and LUSC. Distance between the centroids was empirically evaluated and visualized using 1 minus the Pearson correlation coefficient for the centroids (Figure S5.2).

S5.2 Validation of selected genomic alterations of the gene expression subtypes

The key characteristics of expression subtypes from prior work includes: 1) the atypical subtype comprising those tumors with little evidence for chromosome 7 amplification (Figures S5.3) and including essentially all HPV(+) samples (Figure 4); 2) the mesenchymal subtype characterized by expression markers of the epithelial to mesenchymal transformation, including *VIM, DES*, the transcription factor *TWIST1*, and the growth factors *PDFGRA/B* (Figure S5.2); 3) the basal subtype typified by expression patterns seen in the basal layer of the human airway epithelium and overall low levels of the transcription factor *SOX2* relative to *TP63* (Figure S5.4); 4) the classical subtype characterized by the heaviest smoking histories and elevation expression levels of oxidative stress response genes, including the transcription factor *NFE2L2* (Figure S5.2). Each of these characteristics was again observed in the TCGA cohort. Other previously defined characteristics, including an overlap with expression subtypes defined for LUSC (Figures S5.1-2).

S5.3 Subset analyses by genomic platform

Identification of normal samples

RNA-Seq and miRNA-Seq data were available for the same set of 37 paired normal samples, while DNA methylation data were available for 50 paired normal samples. Data from all three platforms were available for a total of 20 paired normal samples. The 37 normal samples with available RNA-seq data were obtained from five different sites: floor of mouth, larynx, oral cavity, tongue, and tongue/base of tongue. This observation, combined with findings obtained from pathology review and unsupervised clustering of the RNA-Seq data (Figure S5.5) suggested the presence of considerable heterogeneity among the 37 normal samples. Therefore a subset of 16 normal samples was identified that were classified as squamous histology and contained at least 30% squamous epithelium. This set of 16 samples was used when performing differential abundance analyses.

mRNA differential abundance analysis

Methods SAMR was used to identify differentially expressed genes between comparison groups of interest, including tumor vs. normal and HPV(+) vs. HPV(-). The results are summarized in Data File S5.1. Gene expression measurements were computed as described in Section S3.1. SAMR[62] was then applied to identify differentially expressed genes using 1000 permutations and an FDR threshold of 0.05.

Results

HPV(+) vs. HPV(-) tumors Genes exhibiting increased expression in HPV(+) samples included transcription factors (*TAF7L*, *RFC4*, *RPA2*, *TFDP2*), DNA replication genes (*MCM2 – MCM7*), and additional cell cycle regulators (*CDKN1B/C*, *CDKN2A/B/D*, *CDC7*), the majority of which were identified in previous studies[63-65]. Lymphocyte markers *CD3D/E* and *CD8A/B* exhibited increased expression in HPV(+) samples, as did *E2F1-E2F4*, DNA polymerases, and *XRCC1*. *NAP1L2* and *KIRREL* were more lowly expressed in HPV(+) tumors, as noted previously[63].

HPV(+) vs. HPV (-) in oropharynx tumors Because HPV(+) tumors occur primarily in the oropharynx, tumor site may be a confounding variable when assessing differences in expression patterns between HPV(+) and HPV(-) tumors. Therefore we repeated the above analysis after restricting to oropharynx tumors. Of the genes noted above, *CDKN2A/C*, *RPA2*, *MCM2/3/5*, and *TAF7L* were up-regulated.

Larynx tumors vs. other sites We detected increased expression of oxidative stress response genes *AKR1C1/2/4*, which have been associated with chemoresistance[66]. Elevated expression of stem cell marker *ALDH1A1* was noted[67], as was increased expression of other aldehyde dehydrogenase genes. *FGF2/3* and *FGFR2* exhibited increased expression[68], as did additional genes up-regulated in larynx tumors including hedgehog signaling pathway genes *GLI1/2* and *PTCH1*[69] and the transcription factor and stem cell marker *SOX2*[70].

Oral cavity tumors vs. other sites Increased expression of several S100 family members, including *S100A7/8/9* was noted, as has been seen in other tumor types[71]. We also observed elevated expression of stratified epithelial markers *KRT14/17*, hyperproliferative keratinocyte marker *KRT6*[72], and *CD44*, a recognized proliferation and cancer stem cell marker[73]. Genes exhibiting lower expression included *ATR*, *BRCA1/2*, *FANCA/B/C/D2, PPARG, TP53, VHL,* and *XPC*, several of which are associated with DNA repair. Notably, Sparano et al.[74] found that many of these genes exhibited chromosomal loss in their study of oral squamous cell carcinoma.

Oropharynx tumors vs. other sites Most up- or down-regulated genes in oropharynx tumors showed similar expression patterns in HPV(+) tumors. Of those genes noted above, *TAF7L*, *RFC4*, *RPA2*, *MCM2/3/5/6/7*, *CDKN1B, CDKN2A, CDC7, CD3D/E, CD8A/B* had elevated expression in oropharynx tumors, while *KIRREL* had lower expression.

Tumor vs. normal The meta-analysis of Yu et al. lists 25 genes that exhibited increased and decreased expression in 41 tumor vs. normal studies of HNSCC, almost all of which were detected here. Among these are *COL1A2/4A1/5A2, FN1, IL8, KRT17, MMP1/3/10/12*, which were up-regulated in tumors, and *ECM1, EMP1*, and *KRT4/5/13*, which were down-regulated in tumors. A number of additional collagen genes not mentioned in Yu et al.[75], were up-regulated, as were *HLA-A/B/C*.

miRNA differential abundance analysis

Methods

We used SAMseq's (samr v2.0, R 3.0.2)[62] two-class unpaired analyses with a read count input matrix and an FDR threshold of 0.05 to identify miRs that were differentially expressed (see Overview in Data File S5.2). Each run generated a pair of files: genes "up" and "down". We filtered each file by removing miRs with median expression less than 50 RPM in either of the input sample groups, and miRs for which the Wilcoxon adjusted p-value was greater than 0.05; then ranked the filtered results by a median-based fold change.

Results

Differential abundance analyses between tumor and adjacent normal samples identified up and down-regulation of several previously reported oncomirs and tumor suppressor miRNAs, respectively (Figure S5.6). These included miR-21, one of the most well characterized oncomirs across human cancers [73,76], as well as miR-375[77-79], miR-196a [80,81], miR-193b [82], and miR-99a [83], which have all been previously found to be aberrantly expressed in HNSCCs.

Analyses comparing HPV(+) vs. HPV(-) samples identified several miRNAs that have been reported as correlated with HPV(+) status in HNSCC. In particular, the two most up-regulated (miR-9-5p, miR-20b-5p) and the most down-regulated (miR-193b-3p) miRNAs in HPV(+) oropharyngeal cancers (OPCs) (Figure S5.7b) were similarly associated with OPC HPV status in previous work[14,15,17].

For comparisons by anatomic site, both one-vs-all and one-vs-one analyses identified a number of miRNAs that were differentially abundant (Figure S5.8). For example, miR-150-5p was consistently more abundant in oropharyngeal (OP) tumors than in oral cavity (OC) and laryngeal tumors (fold change (FC)≈2.7, Figure S5.8 b,e,f). This miR has been reported as one of the most up-regulated in OP compared with OC tumors[17]. While the miR's differential abundance analysis based on HPV status across all anatomical sites suggested that its abundance may be associated with HPV(+) tumors (FC=2.6, Figure S5.7a), for OP samples the miR was not differentially abundant between HPV(+) and HPV(-) tumors at an FDR of 0.05 (Figure S5.7b), suggesting that high miR-150-5p abundance is related more to the OP anatomical site than to HPV status. In contrast, the HPV(+) vs. HPV(-) fold change for miR-9-5p increased from 21.9 for samples at all sites to 27.0 for OP samples (Figure S5.7), suggesting that the high abundance of miR-9-5p observed in OP samples compared with other sites (Figure S5.7, Figure S5.8) is driven primarily by HPV infection.

Epigenetically silenced genes

To identify epigenetically silenced genes we implemented a modified version of the criteria described previously[26]. Specifically, we first identify promotor CpG sites that meet several criteria: (a) at least 95% of normal samples should be clearly unmethylated ($\beta <= 0.1$) at that site, (b) at least 5% of tumor samples should be clearly methylated ($\beta >= 0.3$) and (c) a t-test comparing expression levels in methylated ($\beta >= 0.3$) and unmethylated tumor samples ($\beta <0.1$) should be significant at an FDR < 0.01. A gene is defined as epigenetically silenced if at least 25% of the promoter CpG sites meet all of these criteria. Applying this procedure to 279 tumor and six normal squamous mucosa samples yielded a set of 847 epigenetically silenced genes shown in Data File S5.3.

Differential DNA methylation between tumor sites, HPV(+) smokers and non-smokers, HPV(+) and HPV(-) samples, and oropharynx only HPV(+) and HPV(-) samples

We used empirical Bayes modified t-test from the limma package [18] in R to find probes that were differentially methylated between different tumor sites, as well as between HPV(+) smokers and non-smokers, HPV(+) and HPV(-) samples, and oropharynx only HPV(+) and HPV(-) samples. Results of all pair-wise comparisons are shown in Data File S5.4.

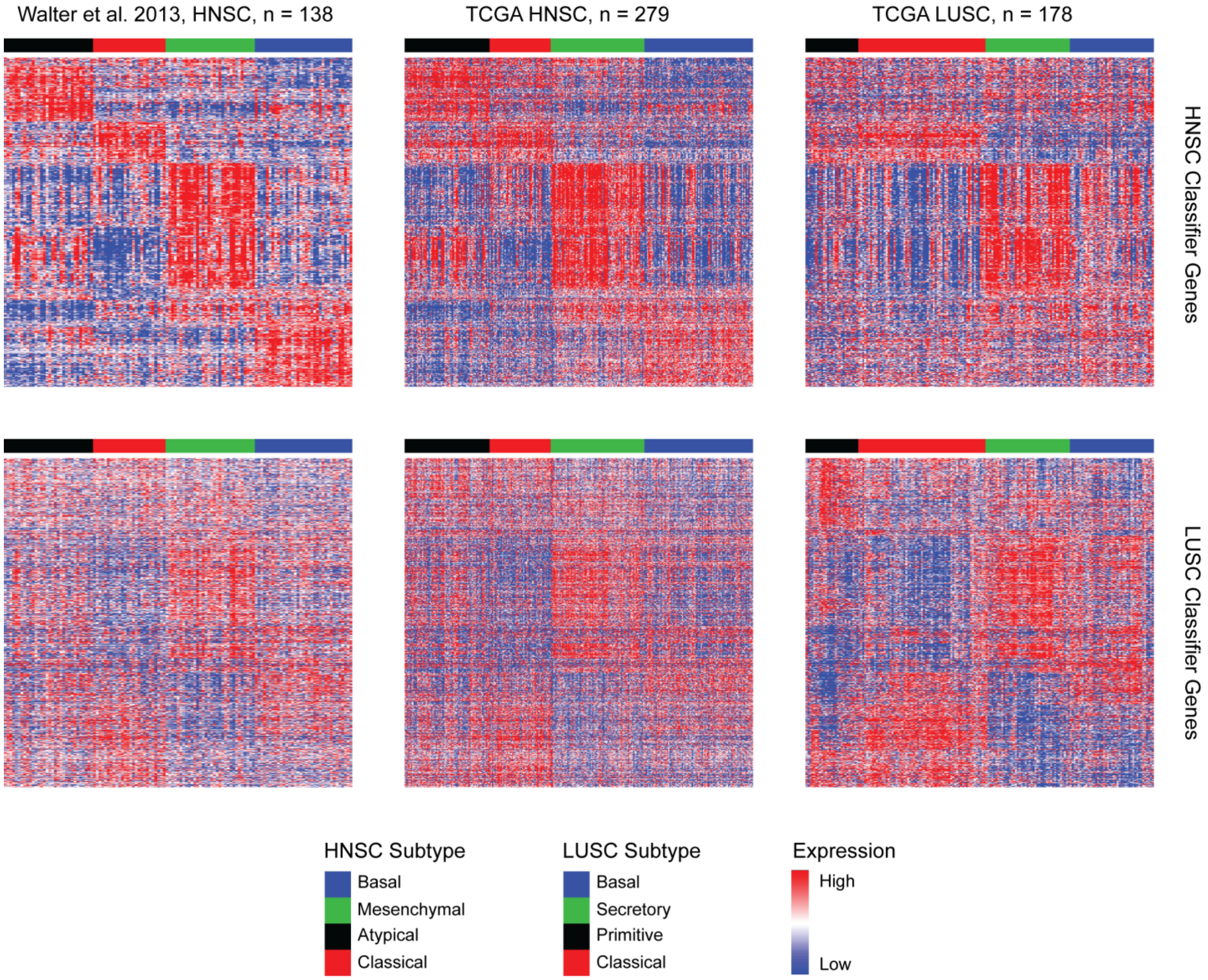


Figure S5.1 Comparison of gene expression patterns in squamous cell carcinomas of the upper aerodigestive tract. Three datasets are displayed in the panels and named as the panel header (Walter et al., 2013; TCGA HNSC, TCGA head and neck squamous cell cohort; and TCGA LUSC, TCGA lung squamous cell cohort). The top row of panels displays a set of 728 genes discussed in S5.1 which constitute a “subtype predictor.” Genes are ordered according to their predicted classes from Walter and colleagues[60]. Samples are ordered by their predicted class. This method documents that gene and sample ordering from the Walter et al. 2013 [60] dataset recapitulates clusters in the TCGA HNSC dataset. In the final panel of the top row, the gene order remains the same, but in place of predicted classes from HNSC, samples are named and ordered according to gene expression subtypes previously defined for lung squamous cell carcinoma (primitive, classical, secretory, and basal) as described in the 2012 TCGA LUSC report [31]. In the bottom row of panels, the genes predicting LUSC expression subtypes are displayed in the order defined by the LUSC predictor. The sample ordering in the bottom row of panels is identical to that from the top row. Ordering samples and genes in this manner demonstrates shared patterns across all three datasets and squamous cell carcinomas from different organ sites.

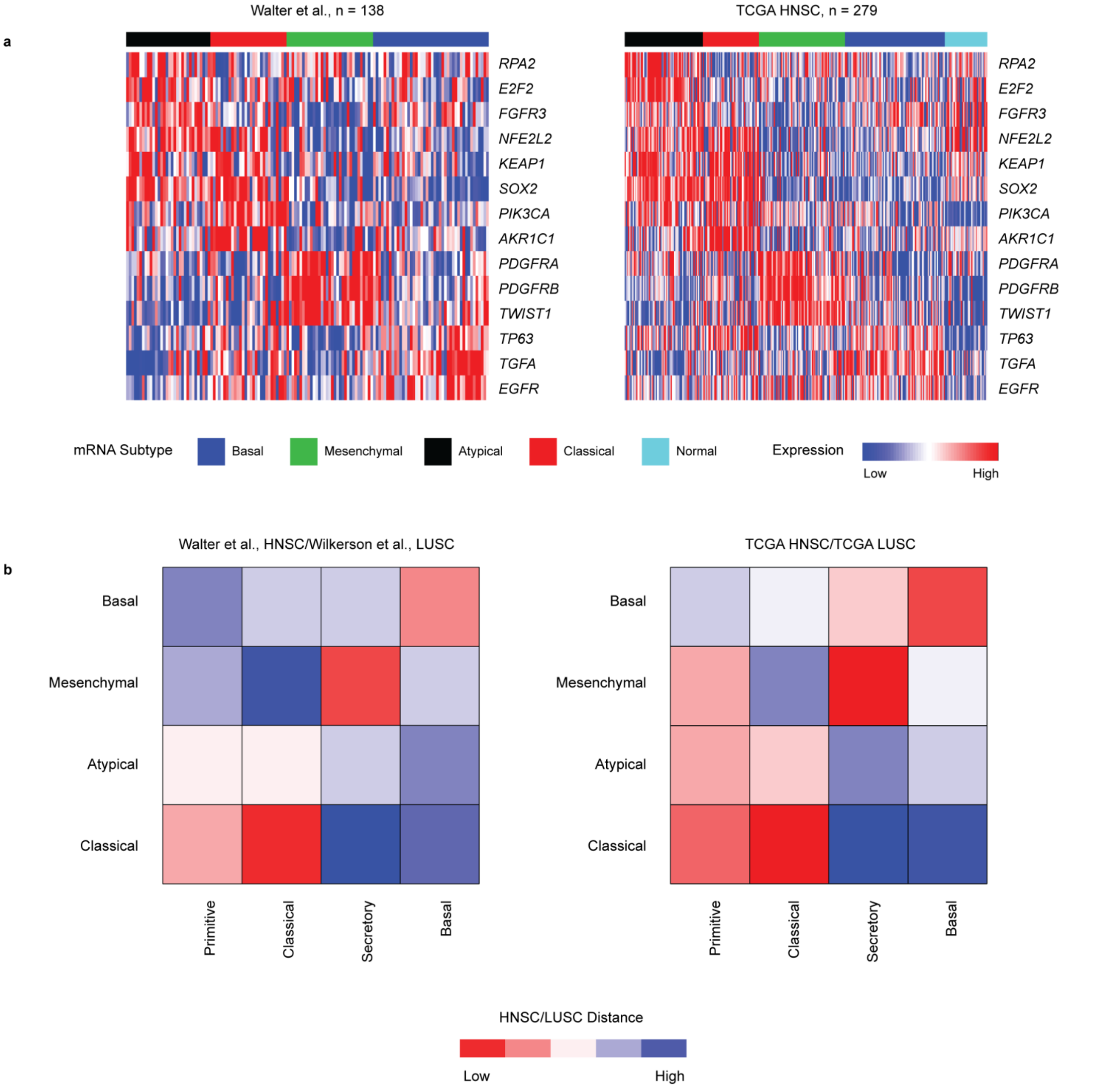


Figure S5.2. Comparison of select genes and expression subtype centroids for squamous cell carcinomas of the upper aerodigestive tract. (a) Gene expression patterns for select genes are displayed by heatmaps in which rows are genes and columns are tumor samples grouped by expression subtype. Expression levels are indicated by the shading given in the heatmap. (b) Heatmap display of correlation-based distances between centroids for expression subtypes in HNSC (rows) and LUSC (columns). Comparisons are shown for the Walter et al. [60] and Wilkerson et al. [84] (left) and TCGA HNSC and LUSC (right) cohorts.

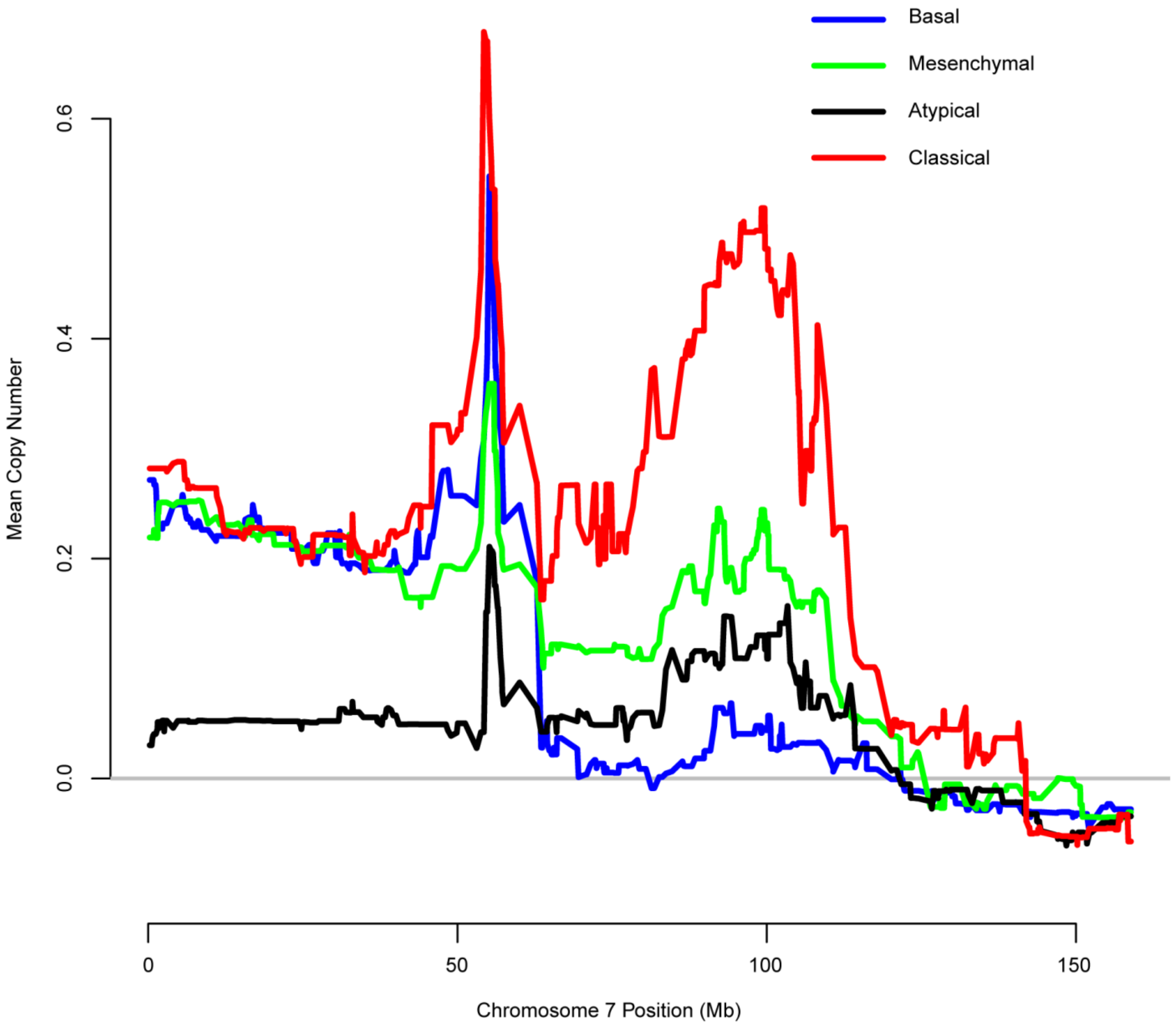


Figure S5.3. DNA copy number in chromosome 7 by gene expression subtype. For each gene in chromosome 7 the mean of the segmented GISTIC copy number values, as described in Section S2.1, in each expression subtype was computed and plotted in genomic order. Focal amplifications of the *EGFR* locus (55.2 Mb) are comparatively rare in the atypical subtype.

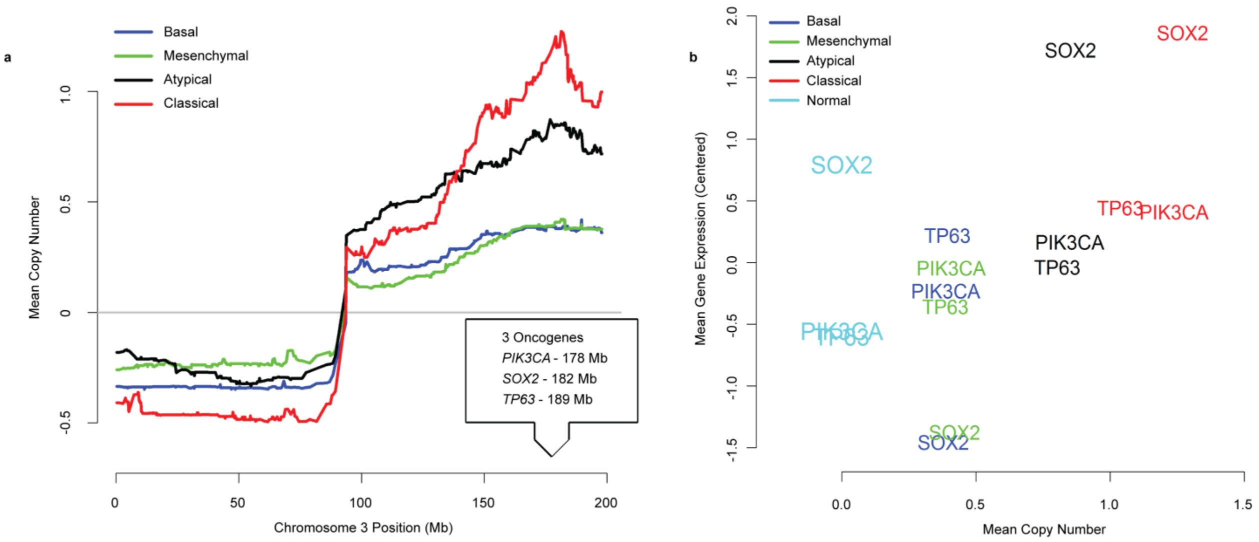


Figure S5.4. DNA copy number and gene expression of canonical oncogenes in chromosome 3q by gene expression subtype. (a) Mean copy number by subtype along chromosome 3, documenting near universal deletions of 3p and amplifications of 3q, although the relative peak heights vary by tumor subtype. (b) Mean DNA copy number and gene expression values for candidate oncogenes *TP63*, *PIK3CA*, and *SOX2* [70] are shown for each of the gene expression subtypes and normal tissue (n = 16). Mean DNA copy number values were computed using the segmented GISTIC copy number data described in S2.1. Normal samples were assigned to have segmented copy number equal to zero consistent with an intact genome. Mean gene expression values were computed after combining expression measurements from tumor and normal samples and median centering by gene.

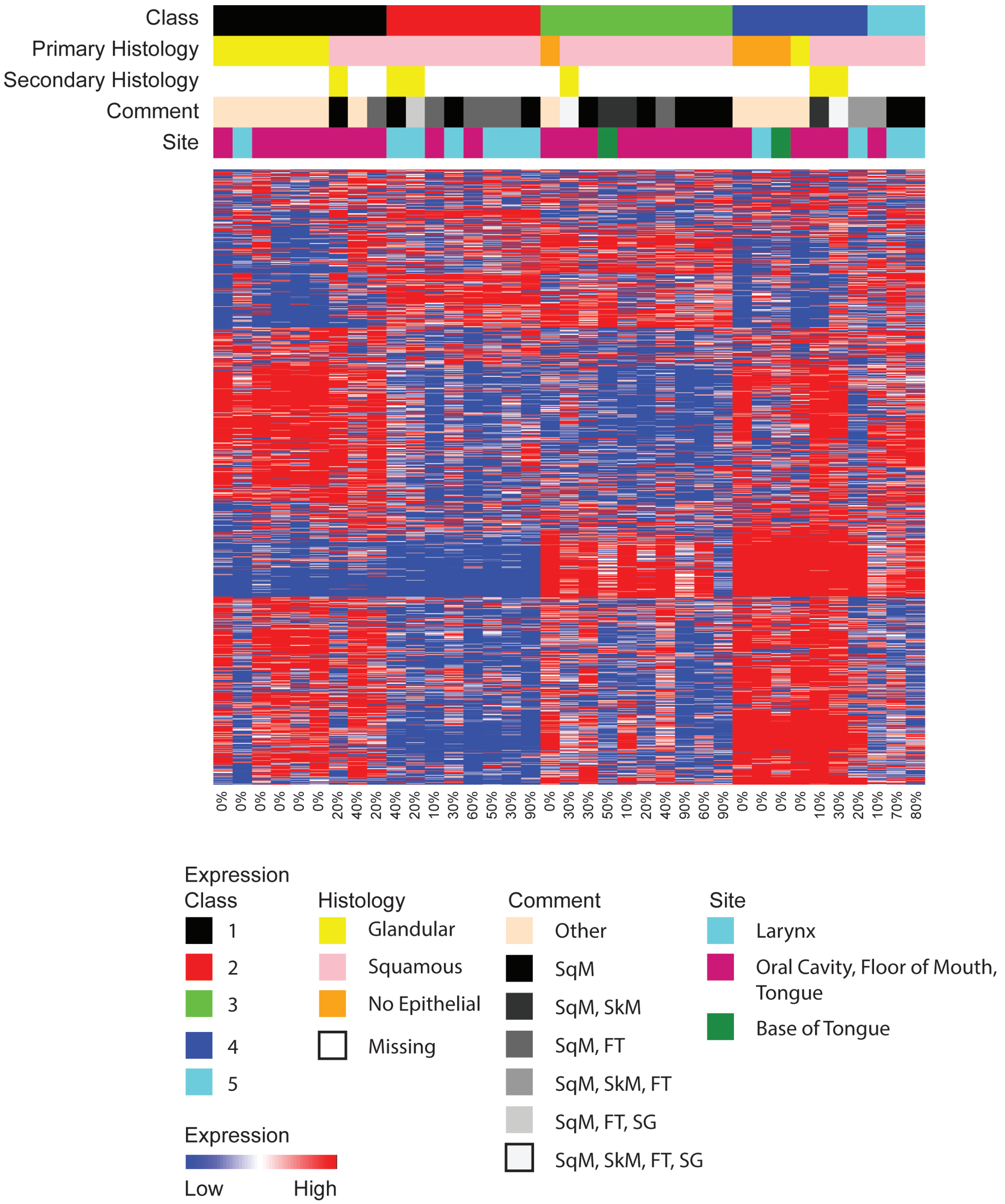


Figure S5.5. Gene expression heatmap for 37 normal samples. The 2500 most variably expressed genes appear in rows and samples appear in columns. Unsupervised clustering identified multiple subsets of normal samples exhibiting distinct gene expression patterns. Samples are ordered according to predicted subtype, then primary histological classification. The percentage of squamous epithelium in each sample is indicated at the bottom of the heatmap. SqM = squamous mucosa; SqM, SkM = squamous mucosa and skeletal muscle; SqM, FT = squamous mucosa and underlying fibrous tissue; SqM, SkM, FT = squamous mucosa with underlying

skeletal muscle and fibrous tissue; SqM, FT, SG = squamous mucosa with underlying fibrous tissue and seromucinous glands; SqM, SkM, FT, SG = squamous mucosa with underlying skeletal muscle, fibrous tissue, and seromucinous glands.

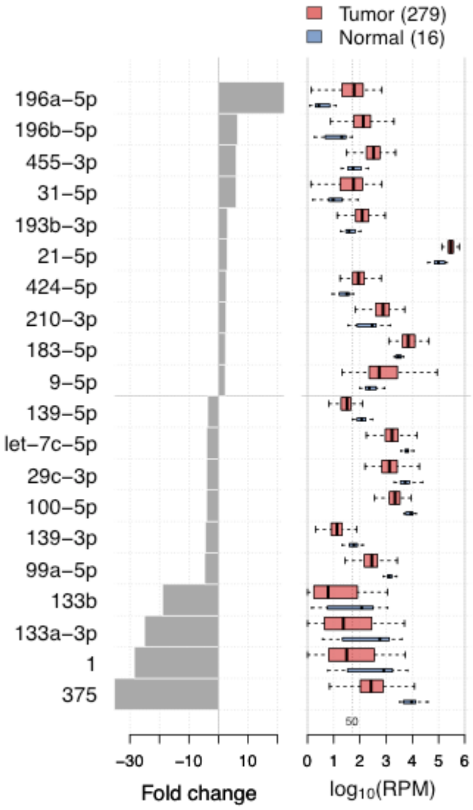


Figure S5.6. miRs that are differentially abundant between tumor and adjacent normal samples. Results are from an unpaired analysis, N(normals, tumors)=16, 279. Left: fold change, linear scale. Right: distributions of RPM abundance, log scale. Up to 10 of the largest fold changes in each direction are shown; FDR ≤ 0.05 (see S5.3). Blue triangles mark miRs that are known to be associated with EMT.

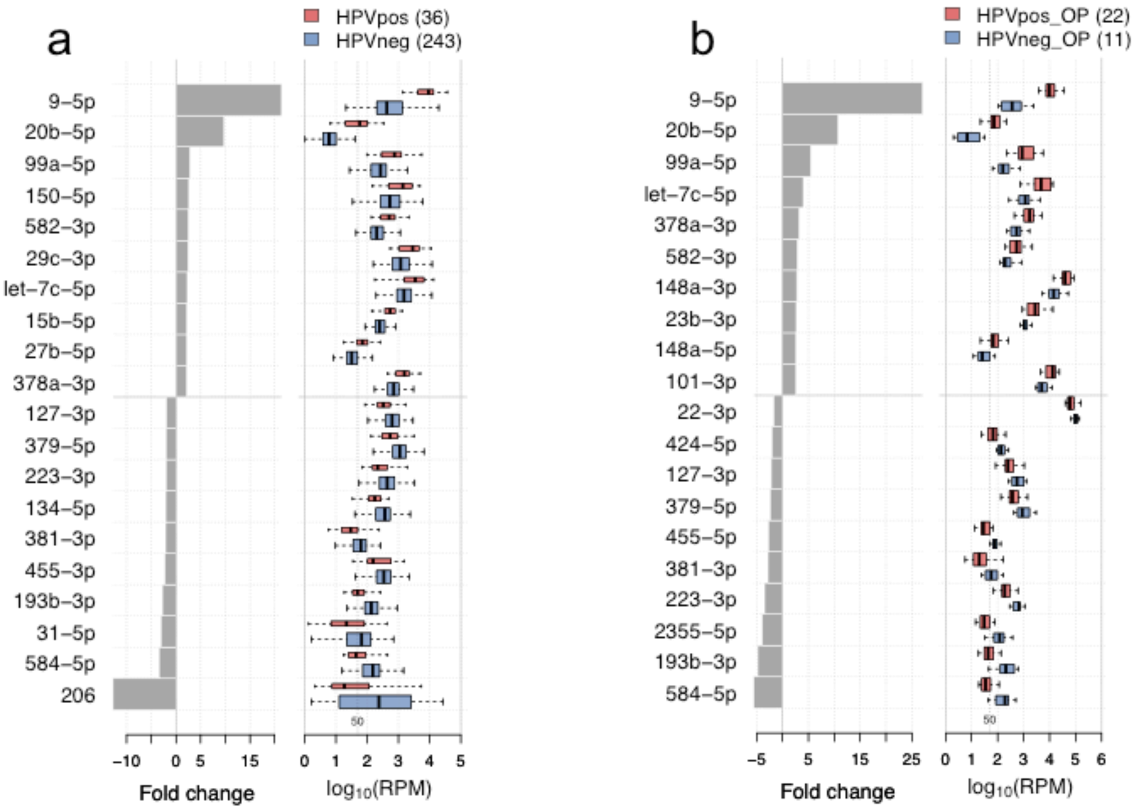


Figure S5.7. miRs that are differentially abundant between HPV(+) and HPV(-) samples. a) All 279 samples, HPVpos (N=36) and HPVneg (N=243). b) Only the 33 oropharynx (OP) samples, HPVpos_OP (N=22) and HPVneg_OP (N=11). Left: median-based fold change, linear scale. Right: distributions of RPM abundance, log scale, with black vertical lines showing medians. Up to 10 of the largest fold changes in each direction are shown; FDR ≤ 0.05 (see S5.3). HPVpos, HPV positive; HPVneg, HPV negative.

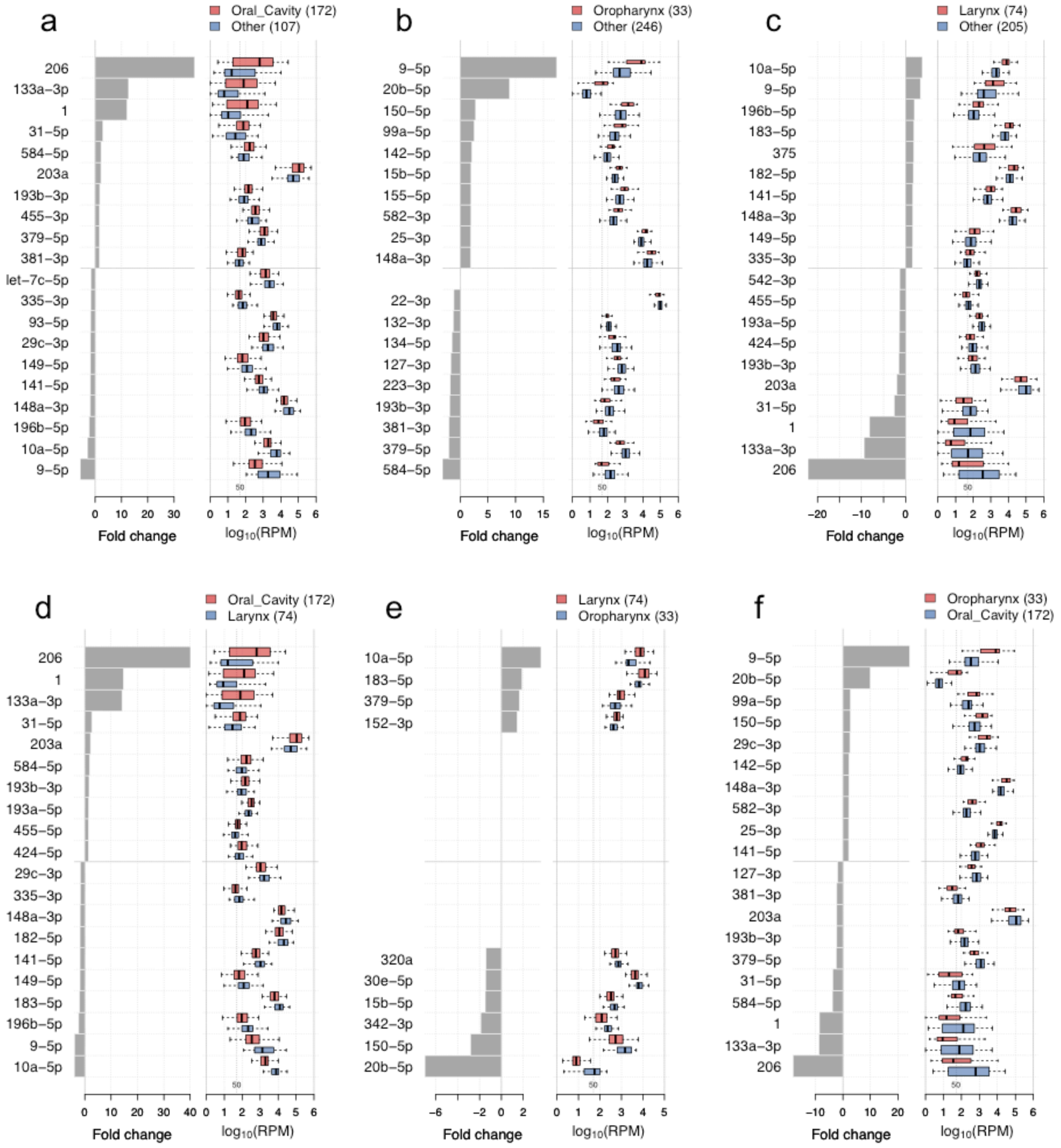


Figure S5.8. miRs that are differentially abundant between different anatomic sites. Oral cavity (N=172), oropharynx (N=33), larynx (N=74). a-c) One site vs. all other samples. d-f) One site vs. one site. Left: median-based fold change, linear scale. Right: distributions of RPM abundance, log scale, boxplots with medians. Up to 10 of the largest fold changes in each direction are shown; FDR ≤ 0.05 (see S5.3).

S6: Reverse phase protein array analysis

S6.1 Methods and statistical analysis

Protein lysate was prepared and analyzed by reverse phase protein array (RPPA) as previously described [85-89]. Briefly, protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 nmol/L Hepes (pH 7.4), 150 nmol/L NaCl, 1.5 nmol/L MgCl2, 1 mmol/L EGTA, 100 nmol/L NaF, 10 nmol/L NaPPi, 10% glycerol, 1 nmol/L phenylmethylsulfonyl fluoride, 1 nmol/L Na3VO4, and aprotinin 10 Ag/mL) from human tumors and RPPA was performed. Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration and boiled with 1% SDS. Tumor lysates were manually diluted in five-fold serial dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 160 primary antibodies (Data File S6.1) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction.

Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI [87,89], available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the y-axis and the relative log2 concentration of each protein on the x-axis using the non-parametric, monotone increasing B-spline model [85]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric [89] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data, https://tcga-data.nci.nih.gov/tcga/tcgaAbout.jsp). Protein measurements were corrected for loading as described [87,89,90] using median centering across antibodies (level 3 data, https://tcga-data.nci.nih.gov/tcga/tcgaAbout.jsp). In total, 160 antibodies and 200 samples were used in the analysis (of 279 in the core sample set, Data File S6.2). Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described [91]. These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described [91]. Raw data (level 1, https://tcga-data.nci.nih.gov/tcga/tcgaAbout.jsp), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC. Differences in protein expression in marker positive and negative groups (e.g. mutation vs. wildtype tumors, HPV(+)/(-) (Figure S6.1), and amplified/non-amplified (Figure S6.2)) were determined by t-test.

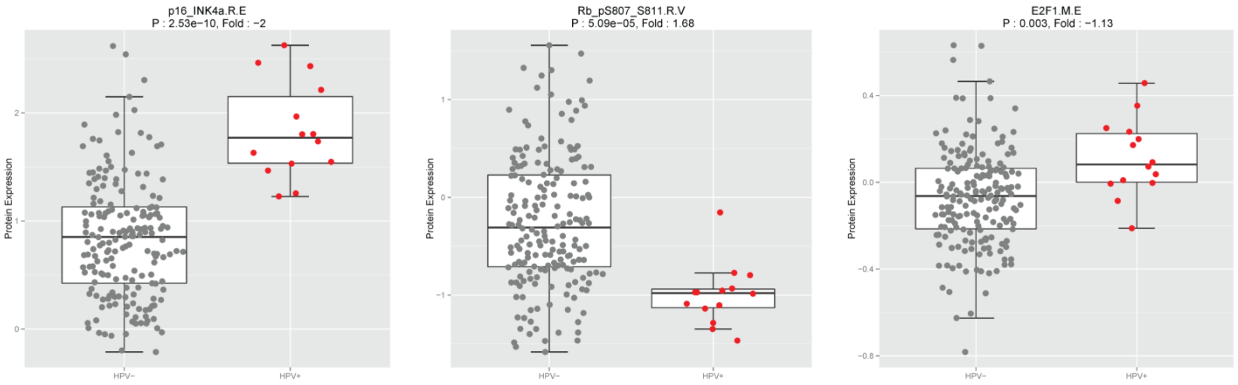


Figure S6.1. Protein expression of p16, pRb, and E2F1 by HPV status. Protein profiling demonstrates increased expression of p16 and E2F1 and decreased expression of pRb in HPV(+) tumors.

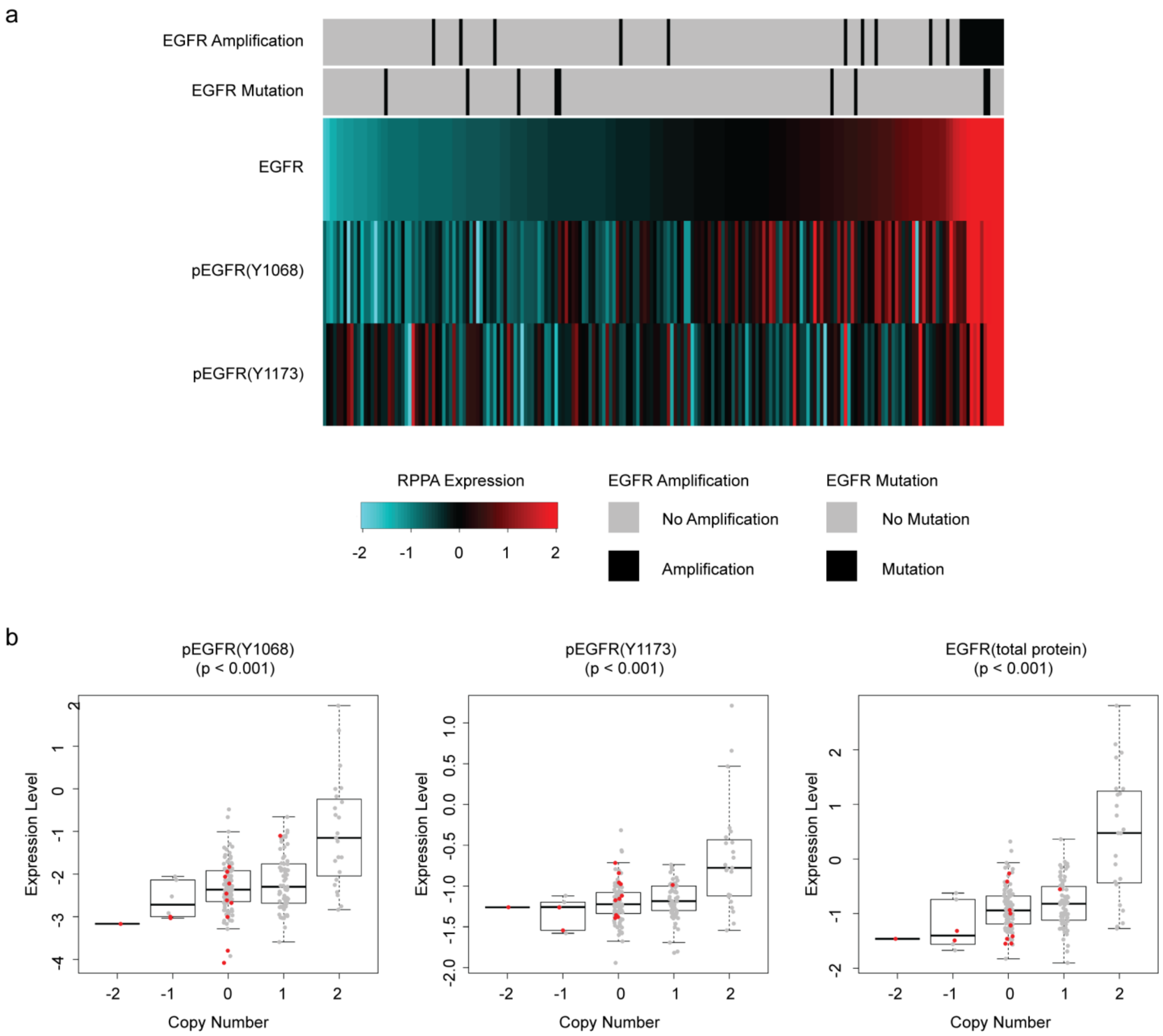


Figure S6.2. RPPA analysis of EGFR as a function of *EGFR* amplification. (a) Amplification in *EGFR* gene is most strongly correlated with increased total and phospho-EGFR protein levels (by t-test) out of 160 proteins measured by RPPA. (b) Changes in *EGFR* copy number are also directly correlated with protein expression levels (by Spearman rank correlation, corresponding p-values shown). HPV(+) tumors (red) did not carry EGFR amplifications and expressed lower levels of total (p=0.027) and phospho-EGFR (p=0.016) levels as compared to HPV(-) tumors (by t-test between HPV(+) and HPV(-) tumors.

S7: Pathways and integrated analysis

S7.1 MEMo analysis of co-occurring and mutually exclusive genomic events

As shown in Ciriello et al. [92], certain signaling pathways may be disrupted by genomic events involving several different constituent genes – e.g. homozygous loss of *CDKN2A*, amplification of *MDM2/4*, and mutation of *TP53* may all disrupt the p53 signaling pathway. As a result, these genomic events may exhibit evidence of mutual exclusivity when multiple tumor samples are examined. Conversely, certain genomic events – e.g. copy number gains or losses – may affect a number of genes, and these co-occurring alteration patterns may be seen across multiple tumor samples. The Mutual Exclusivity Modules [92] (MEMo) component of the cBioPortal [56,57] was used to assess the statistical significance of the observed mutual exclusivity or co-occurrence of select genomic events (Figure S7.1, Data File S7.1). Notable findings include the mutual exclusivity of *CASP8* mutations and amplifications of *FADD*, as well as the co-occurrence of *CASP8* and *HRAS* mutations, as shown in Figures 4 and S7.1.

In some cases, statistically significant associations of gene alterations might be attributable to physical proximity in the genome, such as the highly statistically correlated amplifications of *TP63*, *PIK3CA*, and *SOX2* on chromosome 3q. Regions such as 3q which harbor multiple candidate driver genes in close genomic location can be difficult to interpret as these genes are generally co-amplified or co-deleted such that the relative importance of each gene is not independently assessable. Integrated genomic analysis, however, suggests some potentially differential behavior of genes within the locus such as 3q across the sample set, such as was seen for gene expression of *SOX2* versus *TP63* in 3q as a function of expression subtype in Figure S5.4. MEMo analysis suggests a second example in which genes from 3q may play a different role in different tumors. While *TP63* and *SOX2* were statistically associated with amplification of chromosome 11q13, *PIK3CA* alteration showed no association. By contrast, *PIK3CA* alteration was correlated with amplification of 11q22 *YAP1* locus, whereas there was no association for *SOX2* or *TP63*. Activation of the PI3K-AKT pathway has previously been linked mechanistically to *YAP1* protein phosphorylation, cytoplasmic localization, and inactivation in an HNSCC lacking *TP63* overexpression, consistent with the genetic linkage between these events [93].

S7.2 Genomic aberrations in gene expression subtypes

For the purpose of displaying the coordinated statistical relationship between molecular lesions of interest, an additional set of analyses was performed, similar to that shown in Figure 4. Gene mutation status was based on the results of exome sequencing, as described in Section S4. Copy number gains or losses were assessed using the GISTIC discrete copy number values as described in Section S2.1. Binary methylation data for *CDKN2A* were computed using the procedure described in Section S8.1. Discrete gene expression values were calculated by standardizing gene expression measurements by gene and binning. Gene inactivation is defined by the presence of gene mutation, homozygous loss, gene methylation, or low expression; gene activation is defined by gene mutation, high copy number gain, or high expression.

DNA copy number and gene expression in chromosome 11q

Multiple amplification and deletion peaks were selected as significantly altered by copy number analysis, and these were further highlighted by their statistical significance in the MEMo analysis described above. We therefore displayed these both in terms of chromosomal location and by gene expression subtyping (Figure S7.2). By chromosome analysis the pattern of 11q13 amplification was clearly seen, almost always in conjunction with telomeric loss and occasionally with 11q22 gain as well. The 11q22 gain was rarely seen without the 11q13 gain. The finding of a co-amplification suggested the possibility of a structural rearrangement in which 11q13 and 11q22 might be rearranged in spatial proximity in tumors harboring the combined event. We therefore queried the low pass WGS data and RNA-Seq data for evidence of alignments with split reads involving 11q13 and 11q22. No evidence was found for split reads (one mated read mapping to 11q13 and 11q22) supporting a structural rearrangement bringing these two regions in close proximity. We therefore concluded that there was no support for a single structural rearrangement as the source of the co-amplification of these two loci. It is noteworthy that loss of 11q22 is seen in most HPV(+) tumors. Coordination with genes relevant to the amplifications (*CASP8*, *BIRC2*, *YAP1*, *CCND1*, *FADD*, and *HRAS*) is discussed in the main text.

When viewed in light of gene expression subtypes, the integration of expression of target genes from the region may reveal additional insights. Amplifications of 11q13 are seen in all 4 tumor subtypes in association with coordinated overexpression of three putative target genes, *CCND1*, *FADD*, and *CTTN*. However, the co-amplification of 11q22 in association with preserved overexpression of 11q13 target genes and 11q22 target genes was essentially limited to the basal subtype. Also nearly unique to the basal tumors is an additional event of *ATM* deletion and relatively low expression. *ATM* is located approximately six megabases telomeric of the 11q22 amplification peak, offering support that it may be one of the targets of the deletion peak annotated to 11q22. *ATM* is not included in the deletion peak, although it is adjacent and the data here suggest it is a credible candidate.

DNA copy number and gene expression for HLA class 1 and lymphocyte signature genes

A role for somatic mutation in the cancer hallmark of immune escape/selection was suggested by the presence of inactivating mutations in the *HLA-A* gene and other related pathway alterations, which were seen primarily in the basal subtype (Figure S7.3). Similar somatic loss-of-function alterations of *HLA-A* were reported recently in genomic studies of lung cancer [31]. We therefore investigated the association between *HLA-A* mutations, deletions, gene expression, and selected gene expression markers of inflammation as a function of molecular subtype. Interestingly, the mesenchymal subtype was notable for high levels of innate immunity, in particular natural killer (NK) cell marker *CD56* expression and low frequency of HLA class I mutations, which would abrogate HLA class I expression and trigger NK cell elimination of tumor cells. Given the recently reported efficacy of anti-Programmed Death 1 (*PD1*) and anti-Cytotoxic T-Lymphocyte Antigen 4

(*CTLA4*) antibodies in non-small cell lung cancer (NSCLC) [94,95], which demonstrated similar genomic and expression alterations to HNSCC, *HLA-A* mutations suggest potential implications for success of these targeted antibodies, as well as biomarkers of successful T cell based immunotherapy.

S7.3 Exploratory clustering / Unsupervised analysis of genomic platforms

Previous genome-wide profiling studies of HNSCC [78,96,97] have been either relatively small or limited to a single platform. Therefore the size of the cohort and the diversity of available genomic data provides unprecedented opportunity to perform an integrated genomic analysis of the disease. To extend the ability to associate patterns in genomic data, exploratory clusters were generated from reverse phase protein, miRNA sequencing, DNA methylation profiling, and mRNA sequencing data.

Reverse phase protein array clustering

Unsupervised clustering of reverse phase protein array data was performed by non-negative matrix factorization (NMF) clustering, identifying 4 subgroups of tumors with distinct protein expression patterns (Figure S7.4a). RPPA subgroups were significantly associated with subtypes identified in independent data types (including mRNA clustering and methylation subtypes) and with clinical characteristics (tumor site, stage, and grade). It is important to recall that a direct comparison to the entire data freeze sample set of 279 patients was not possible because samples were not available for every case. It is equally important to point out that samples were not missing at random, with a greater proportion of oropharynx samples missing due to the fact that these samples were generally smaller and less likely to provide tissue for RPPA analysis. Such a systematic bias significantly hinders the ability to comment on the integrated signatures of RPPA as they relate to HPV. Despite these limitations, statistically significant association was seen between RPPA subtypes and subtypes from other platforms, including one subgroup defined by a prominent previously reported EMT pathway signature (Figure S7.4b) [98]. Correlation of RPPA subtypes identified 20 genes with mutations with high correlation to RPPA subtypes are shown (Figure S7.5, $p < 0.05$) although none of these were found on the statistically significantly mutated gene list.

miRNA clustering

Unsupervised NMF consensus clustering using 304 of the most variant 5p-3p-strand microRNA sequence (miRNA-Seq) data for 279 tumor specimens suggested a five-group solution using methods previously described (Figure S7.6) [99]. Distributions of sequencing metrics, tissue source sites, biospecimen core repository (BCR) batches, and purity suggested that the data contained no strong technical biases. A heatmap was generated for the 32 discriminatory miRNAs that had the top 5% of scores in each of the 5 NMF metagenes [100]. For the heatmap, columns (samples) in a reads-per-million (RPM)-normalized abundance matrix were ordered to match the NMF result, the abundance matrix for matched tissue normals was added, then rows (miRNA abundances) were log2-transformed, mean-centered, and were hierarchically clustered using an absolute centered correlation distance metric and average linkage [101,102]. 5p and 3p strand names were assigned using miRBase v19. Purity and ploidy were calculated by the Broad Institute using the method reported by Carter and colleagues [29]. Differentially abundant miRNA 5p and 3p strands (FDR < 0.1%) were identified with RPM-normalized data and SAMseq v2.0/R 3.0.1. Relevant to the EMT pathway signature above, it was noted that in 2 of the 5 groups miR-200 family members were less abundant and EMT scores were higher (Figure S7.7). Per sample EMT scores were then calculated as reported by Byers et al. [98]

DNA methylation clustering

DNA methylation clusters were based on CpG sites with sample to sample variation in the top 1% of all probes. Consensus clustering was applied as implemented in the Bioconductor package ConsensusClusterPlus [103], with Euclidean distance and partitioning around medoids (pam) used to derive clusters (Figure 4b, and Figure S7.8). Student's t-tests were used to test for association between mutations and DNA methylation patterns and Fisher's exact test to test for associations between methylation subtype and other molecular factors including mRNA and miRNA expression and the EMT signature.

In addition to remarkable correlation between mutations of *NSD1* and *NOTCH1* and the DNA hypomethylated cluster, as noted in Figure 4b, the methylation clusters are significantly associated with mutations in *KEAP1*, *SMARCA4*, and *MLL2* (Figure S7.8), although none of these is associated with a unique DNA methylation phenotype. *SMARCA4* is also statistically associated with RPPA clustering above. A number of additional molecular associations to the *NSD1*-depleted/hypomethylated cluster become evident with further analysis. These samples also have the highest rates of overall mutation (Figure S7.8b), higher purity (Figure S7.8e), and several miRNAs are significantly differentially expressed in *NSD1*-depleted/hypomethylated samples as well (Table S7.1). Among the most interesting of these are miR200a/b, which are known to suppress the epithelial- mesenchymal transition, a phenotype highlighted by unsupervised clustering by gene expression arrays, RPPA analysis, as well as miRNA above[104]. These genes are highly expressed in the *NSD1*-depleted/hypomethylated samples (Figure S7.8c), suggesting an epithelial state that is confirmed by low values for the RPPA based EMT signature (Figure S7.8d).

mRNA clustering

After removing the RNA-Seq data for the 36 HPV(+) samples, all RSEM values identically equal to zero were replaced by the smallest non-zero RSEM value. A log2 transformation was applied and the values were subsequently median centered by gene. Consensus clustering [103], principal components analysis, and gene expression heatmaps were used to analyze gene expression data from the 2500 most variably expressed genes, as defined using the median absolute deviation. Three gene expression subtypes were identified in the HPV(-) samples, and these were denoted classes 1 (n = 69), 2 (n = 63), and 3 (n = 111). The unsupervised RNA class labels, together with a separate class label for the HPV(+) samples, are shown in the "Unsupervised RNA Subtype" annotation track of Figure S7.9.

The unsupervised RNA subtypes were strongly associated with both tumor site and the RNA subtypes defined in S5.1. Nearly 70% of the class 1 tumors were found in the larynx, and there was considerable overlap between class 1 and both the HPV(-) atypical and classical tumors. Approximately 80% of both class 2 and class 3 tumors were oral cavity. Class 2 samples were found primarily in the basal subtype, while class 3 contained most of the non-class 2 basal tumors and nearly 70% of the mesenchymal tumors.

Cluster of clusters analysis

As described below and shown in Data File S7.2, statistically significant associations were detected between many of the subtypes defined by the various genomic platforms. In an effort to identify these associations we performed unsupervised clustering of the RNA, DNA copy number, DNA methylation, miRNA, and RPPA class labels using a "cluster of clusters" approach similar to the procedure used in the TCGA breast cancer study [99]. The clustering was performed as follows. We began by creating indicator variables of length 279 for each platform-specific subtype. As an example, the indicator variable for the normal-like methylation subtype takes the value 1 if a given subject belongs to that subtype and is 0 otherwise. This produced a binary matrix with 279 columns and 21 rows: four for RNA; three for DNA copy number; four for DNA methylation; five for miRNA; five for RPPA, including one for "RPPA missing."

Because the binary matrix described above contains numerous identical rows and columns, it was not possible to perform consensus clustering directly. Therefore independent normally distributed values with mean 0 and standard deviation 1E-08 were added to each entry. All of these normally distributed values were essentially equal to zero, so any underlying patterns detected in the resulting matrix were identical to those present in the original binary matrix. The ConsensusClusterPlus [103]package was then applied using hierarchical clustering (clusterAlg = "hc"), a distance metric based on the Pearson correlation coefficient (distance = "pearson"), and 80% resampling of both the features and the samples (pFeature = .8, pItem = .8). The graphical output produced by ConsensusClusterPlus strongly suggested the presence of four clusters.

The four "cluster of clusters" (CoC) subtypes described above are denoted CoC classes 1 (n = 87), 2 (n = 61), 3 (n = 76), and 4 (n = 55) and are shown in the "Cluster of clusters" annotation track of Figure S7.9. Even though the CoC class labels were obtained using the subtypes from all platforms, they displayed a remarkable concordance with the RNA classes from S5.1. CoC class 1 was identical to the basal RNA subtype, all 61 members of CoC class 2 belonged to the mesenchymal subtype, CoC class 3 contained 48 of the 49 classical samples, and 53 of the 55 CoC class 4 samples belonged to the atypical subtype. Considering tumor site, CoC classes 1 and 2 were found primarily in the oral cavity, while the majority of larynx tumors were seen in CoC class 3. We observed sizeable overlap between CoC class 3 and both the normal-like methylation class and miRNA class 1. CoC class 4 contained almost all of the HPV(+) samples and was mostly copy number class 3/quiet.

Comparison of subtypes defined by different platforms

Two-way cross tabulations of the class labels for the RNA, miRNA, DNA methylation, DNA copy number, and RPPA subtypes were created using all possible pairwise comparisons, and these are shown in Data File S7.2. RPPA class labels were only available for 200 of the 279 subjects in the data freeze, and samples without RPPA class labels were classified as missing (NA). A Monte Carlo version of Fisher's exact test was used to assess the statistical significance of the association between all pairs of class labels using a total of 10,000 simulations. This analysis was performed (i) using all 279 subjects and grouping all subjects without RPPA class labels into a single class, and (ii) after restricting the class labels from all platforms to the 200 subjects that have RPPA data. The p-values are shown in Data File S7.2. The underlying mRNA, miRNA, and RPPA data can be explored through a compendium of next-generation clustered heat maps (http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/), although it should be noted that the class labels presented in Data File S7.2 are not currently available at this site.

S7.4 Supervised integrated analysis of miRNA, gene expression, and copy number

miRNA and mRNA abundance, and copy number variations (CNV) for 279 tumor and 37 adjacent mucosa specimens were extracted from Level 3 data associated with the data freeze [37]. miRNA read counts for 5p and 3p strands were normalized to RPM (reads per million) aligned to miRBase annotated miRNAs. miRNAs were ranked by RPM variance across the samples, and the most variable 50% with a minimum expression of at least 50 RPM were used for integrated analysis. Gene expression was calculated from RNA-Seq data with RSEM v1.1.13 [34] and zeros replaced with the minimum non-zero RSEM value (0.0033). The most-variant 50% of genes were used for integrated analysis. Both miRNA and mRNA expression data were log2 transformed. The CNV number associated with each gene was defined as the segmented GISTIC [105] value at the corresponding genomic location.

To identify miRNA and mRNA pairs with statistically significant inverse expression relationships, a multi-step approach was implemented. miRNAs were used as independent variables and mRNAs as dependent variables in a linear regression analysis [106-108], which was done for each miR-mRNA pair in the large dataset using the Matrix eQTL [109] R package. To estimate FDR for the regression p-values, we randomly permuted sample labels and compared the observed p-values to the null distribution of p-values obtained by permutation. An FDR threshold of 0.05 was used to identify significant negative miR-mRNA associations. We filtered the predictions from linear regression using experimentally validated miRNA and mRNA interactions from miRTarBase v3.5 [110]. Then, a Mann-Whitney-Wilcoxon test was used to identify differentially expressed miRNAs and mRNAs in tumors and mucosa tissues. Pairs that were regulated in opposite directions were identified, i.e. a down-regulated miRNA and its up-regulated mRNA target, and vice versa. For each miR-mRNA pair, we identified samples that exhibit at least a 2-fold change for both miRNA and mRNA expression (labeled 2FC in Table S7.2), and we used a Fisher exact test to assess whether there was a significant association between miRNA copy number loss and fold change for the miR-mRNA pair. Statistical analyses were performed using R [8] version 2.15.2. Significance was defined as $p < 0.05$.

While a number of miRs were identified by this analysis, two in particular were of note: let-7c and miR-100. In addition to selection by the statistical approach above, we noted that miR-100 happened to fall in a deletion peak (11q23.3) which was statistically significant for HPV(+) but not HPV(-) tumors (although this is the same region harboring *ATM* as discussed above (Figure S7.2) and near a broader deletion peak on 11q23.1 found in HPV(-) tumors as well). As expected, miR-100 was both statistically associated with HPV status, and was associated with deletion and lower gene expression (Table S7.3, Figure S7.10). Let-7c gene expression was statistically associated with copy number but not with HPV status and is not found in a region associated with statistically significant recurrent copy number alteration. For these miRNAs, deletion was associated with increased expression of target genes, including the cell cycle regulator *CDK6*, the transcription factor *E2F1* [111] mitosis regulator *PLK1*, and transcription activator *HMGA2* (Figure S7.10, Tables S7.2, 3) [112-114].

S7.5. Integrated pathway analysis using PARADIGM and PARADIGM-SHIFT

FAT1 alterations are associated with increased Wnt Signaling and significantly co-occur with NOTCH alterations

Recent studies have implicated *FAT1* inactivating mutations and deletions as activators of the Wnt Signaling pathway and promotion of tumorigenesis in multiple cancers, including head and neck cancer [115]. Wildtype *FAT1* was shown to sequester β-catenin, preventing signaling through Wnt which is reversed upon acquiring inactivating mutations to *FAT1*. In order to confirm this finding, we performed differential RNA-Seq analysis looking for enrichment within pathways for samples with either *FAT1* mutations or homozygous deletion. Using genes identified at p-value ≤ 0.01, we looked for pathway enrichment and identified Wnt-signaling as being significantly enriched across multiple pathway databases (KEGG p-value < 0.001, Panther p-value = 0.016). In addition, co-occurrence and mutual exclusivity analysis comparing *FAT1* and members of the pathway associated with squamous differentiation [116] identified a significant co-occurrence of *FAT1* alterations and *NOTCH1* alterations ($p < 0.01$). This corresponds with previous studies that have shown the integrated relationship between Wnt and NOTCH pathways that influence cellular fate [117] and the central role *FAT1* alterations likely play in those pathways.

PARADIGM integrated pathway analysis of copy number and expression data

Integration of copy number, mRNA expression and pathway interaction data was performed on 282 samples using the PARADIGM software [118,119], and this list contains the 279 samples described above. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. RNA-Seq expression and gene copy number data were used as input to the algorithm for this analysis. The mRNA data were converted to relative mRNA expression levels by taking the difference of each gene in each sample to the gene's median computed over 37 normal controls. Data were rank transformed and discretized prior to PARADIGM analysis.

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov and the Reactome database from http://reactome.org. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and are henceforth referred to collectively as pathway concepts. The resulting pathway structure contained a total of 17151 concepts, representing 7111 proteins, 7813 complexes, 1574 families, 52 RNAs, 15 miRNAs and 586 processes.

The PARADIGM algorithm infers an IPL for each gene that reflects a gene's activity in a tumor sample relative to the normal controls. To identify patient subtypes implicated from shared patterns of pathway inference, we ran Consensus Clustering using the top 25% median-centered IPLs ranked by variance implemented on top of the C Clustering Library [101] with 80% subsampling over 1000 iterations of hierarchical clustering based on a Euclidean correlation distance metric (Figure S7.11).

Consensus Clustering of the ~4,300 varying IPLs yielded 8 total clusters which were then filtered down to 4 main PARADIGM clusters with more than 5 samples. These clusters demonstrate distinct pathway activation patterns and significant association with molecular HPV status (chi-squared p-value < 0.0001) and expression clusters (chi-squared p-value < 0.0001). When overlaid with mutation classifications for the 6 genes with mutational frequencies higher than 15%, *TP53* shows significant enrichment (chi-squared $p = 0.0005$) across the clusters.

Pathway-based biomarkers of HPV associated tumors

In order to understand the pathway-level differences of HPV(+) vs HPV(-) tumors, a Wilcoxon rank-sum test was calculated for all ~17K features using molecular HPV status as a group. P-values were adjusted using Bonferroni correction and links were removed from the overall pathway unless both parent and child nodes had a significant p-value. In order to focus the analysis on highly connected features, the parent of each link was required to have at least 10 children overall in the pathway and each child was required to have at least 5 children itself. This resulted in a sub-network of 80 nodes and 105 interactions, visualized using Cytoscape in Figure S7.12. Nodes are colored by their signed Wilcoxon p-value, with significantly higher activity in HPV(+)

samples are colored red , while nodes with significantly higher activity in HPV(-) samples are colored light gray. Of note are the higher inferred levels of E2F transcription factors, *TP53* activity and DNA damage response genes in HPV(+) tumors. In the HPV(-) tumors, increased activity of *SRC*, CAV1, DeltaNp63, and *CCNA1* appear to be the most significant.

To further understand the pathway differences between HPV status and the pathways activated, the set of 25% most variable IPLs was entered into Gene Set Enrichment Analysis (v2.0.13, www.broadinstitute.org/gsea/) along with the set of 833 pathways (out of the original 1,377, filtered by requiring size greater than 5 and less than 1,500). Only two gene sets were found to be significantly enriched at FDR 25%, both positively enriched in HPV(+) tumors. “Cell Cycle G1/S Checkpoint” and “Regulation of Nuclear SMAD2/3 Signaling” both appear to be significantly up-regulated (FDR q-val .112 and .240, respectively). Both of these networks include large numbers of genes identified in the previous sub-network analysis (including *CDKN2A*, E2F, DNA Damage and others).

***NOTCH* and *NFE2L2* mutations are implicated as gain-of-function alterations by evaluating the network neighborhood discrepancies**

In order to assess the functional effect of mutations across the cohort, we employed the PARADIGM-SHIFT algorithm [120] to compare the pathway impact across the set of 15 mutations that were mutated in five or more samples and are annotated in the SuperPathway (*CDKN2A, CASP8, AJUBA, PIK3CA, HLA-A, TGFBR2, HRAS, EPHA2, RB1, KEAP1, RAC1, NOTCH1, TP53, NFE2L2, CUL3*). Mutation neighborhoods were selected in a supervised fashion by selecting features based on the rank ratio of the features determined by a t-test. PARADIGM-SHIFT (P-Shift) scores for the set of mutated genes were computed as the difference in activity between two runs of PARADIGM – one in which only upstream regulators are connected (R-run) and one where only downstream targets are connected (T-run). We then assessed the significance of these scores by comparing them to the distribution of a background model in which the network topology is fixed and the data are permuted without replacement across the set of genes in the neighborhood.

Of the genes tested, *NOTCH1* and *NFE2L2* were found to have significantly better P-Shift scores versus the background model (Z-scores of -2.1 and 5.42, respectively). Figures S7.13 and S7.14 visualize the neighborhoods of these genes using the CircleMap representation [121], where the data across multiple datasets are circularized and sorted by each samples’ shift score within the mutated and non-mutated groups. In the NOTCH analysis, mutations of any NOTCH family gene were considered across the set of samples, and of particular interest are the apparent higher overall activities of *AP1*-related targets *PLAU*, *EDN1* and *TIMP1* in the NOTCH mutated samples. Although the majority of mutated samples appear shifted in the loss-of-function direction consistent with previous reports in head and neck cancer [122], there are a small number of samples that appear as though they may contain gain of function mutations, emphasizing the complex landscape of alterations to this gene. For the *NFE2L2* analysis, mutations included *KEAP1* and *CUL3* mutants, and the P-Shift towards activating indicates mutations to these genes are likely disrupting binding of *NFE2L2* and *KEAP1* preventing inhibition, as has been previously reported[123].

Pathways

Through integrative bioinformatics approaches and manual curation (S7, Figures S7.1-14), we identified a limited number of key pathways to be the targets of high frequency genome alterations in HNSCC overall and in HPV(+/-) subsets (S7.15 parts I and II; Data File S7.3. For HPV(-) tumors, the results underscore the diversity of alterations activating growth factor RTK-RAS-PI3K pathways, and the cell cycle, especially inactivation of *CDKN2A* and *TP53*, that dysregulate control of proliferation and growth[3,9,124-128] . In HPV(+) tumors, activating alterations predominantly involve *PIK3CA, FGFR3* and *E2F1* , while HPV proteins E6/7 provide an alternative mechanism for inactivating non-mutant *TP53* and *RB1* functions in cell cycle control[42,111,128-131]. In a major subset of HPV(-) HNSCC, newly recognized co-amplifications of *FADD* and *BIRC2*, or disruption of *CASP8* in concert with *HRAS* or *PIK3CA* alterations, are components implicated in NF-κB-mediated cell survival and death pathways[124,132-135]. Intriguingly, previously, *TRAF3* inactivation was reported in hematologic malignancies and nasopharyngeal carcinoma [136]. However, here *TRAF3* was inactivated exclusively in HPV(+) tumors, where its disruption could serve a dual role in evading interferon and innate responses to DNA viruses, while enhancing alternative NF-κB pathway activation [135-142]. Loss of the immune recognition determinants *HLA-A/B* and beta 2 microglobulin is supported by protein immunostaining studies[143,144], possibly implicating loss of these genes in escape of both HPV(+) and (-) HNSCC from adaptive immune recognition. Overexpression of the ΔNp63 isoform of *TP63* is observed in both subsets, and linked to inverse regulation of critical NF-κB and *TP53/TP73* gene programs that promote the malignant phenotype [145,146]. ΔNp63 also acts as a key repressor of *NOTCH*, important in epithelial differentiation [9,122,147,148], and this is noteworthy in light of recurrent inactivating mutations in *NOTCH1/2/3* (19%, q < 0.1; 9%, q > 0.1; and 5, q > 0.1, respectively), as well as recurrent mutation and somatic loss events affecting *TP63* the target gene *ZNF750*. Alterations in *KEAP1/CUL3/NFE2L2*, which serve as a sensor for oxidative stress, can promote cancer cell viability[149]. *FAT1* and *AJUBA* are implicated in cytoskeleton and cell polarity[115,150-153]. Together, genome alterations in *TP63, NOTCH, KEAP1/NFE2L2*, or *FAT1/AJUBA*, exhibit links to Wnt-β-catenin signaling[115,147,148,150,154-156], and thus could potentially serve as alternative mechanisms to deregulate normal differentiation. The prevalence of involvement of RTK-PI3K, CCND1-CDKN2A-Let7c-CDK6, FADD/BIRC, NOTCH, and β-catenin pathways suggest the potential for developing therapeutics that target key nodes in these signal pathways. Alterations of *PIK3CA* represent the most common potential target for therapeutics shared by both the HPV(+) and HPV(-) subsets [54,56,57].

S7.6 Somatic alteration in therapeutic targets

Four potentially targetable genetic events occurred with a frequency of greater than10% and showed a differential incidence between HPV(+) and HPV(-) tumors: *EGFR, FGFR, or PIK3CA* mutation or focal amplification and *CCND1* amplification (Figure 3). Although no link between *EGFR* alterations and therapeutic efficacy of EGFR inhibitors in head and neck cancer has been established, it is interesting that the 16% frequency of *EGFR* amplifications and mutations (Figures S4.2) in HPV(-) tumors was similar to the reported efficacy of EGFR inhibitors [157]. Alterations of *PIK3CA* represented the most common potential target for therapeutics shared by both HPV(+) and HPV(-) subsets. *CCND1* was the single most frequently altered oncogene, although it remains to be proven that *CCND1* amplification will serve as a biomarker for responses to the CDK4/6 inhibitors [127]. Fibroblast growth factor receptors (FGFR) were the subject of activating *FGFR3-TACC3* fusion oncogenes in HPV(+) cases and by at least 10% focal amplification and occasional mutations in HPV(-) tumors suggesting a role in a subset of these tumors. Finally, rare aberrations of genes such as *ERBB2, IGF1R, DDR2, HRAS*, and *MYC*, suggests growth factor receptor tyrosine kinases and oncogenes play a role in selected cases, primarily in HPV(-) tumors.

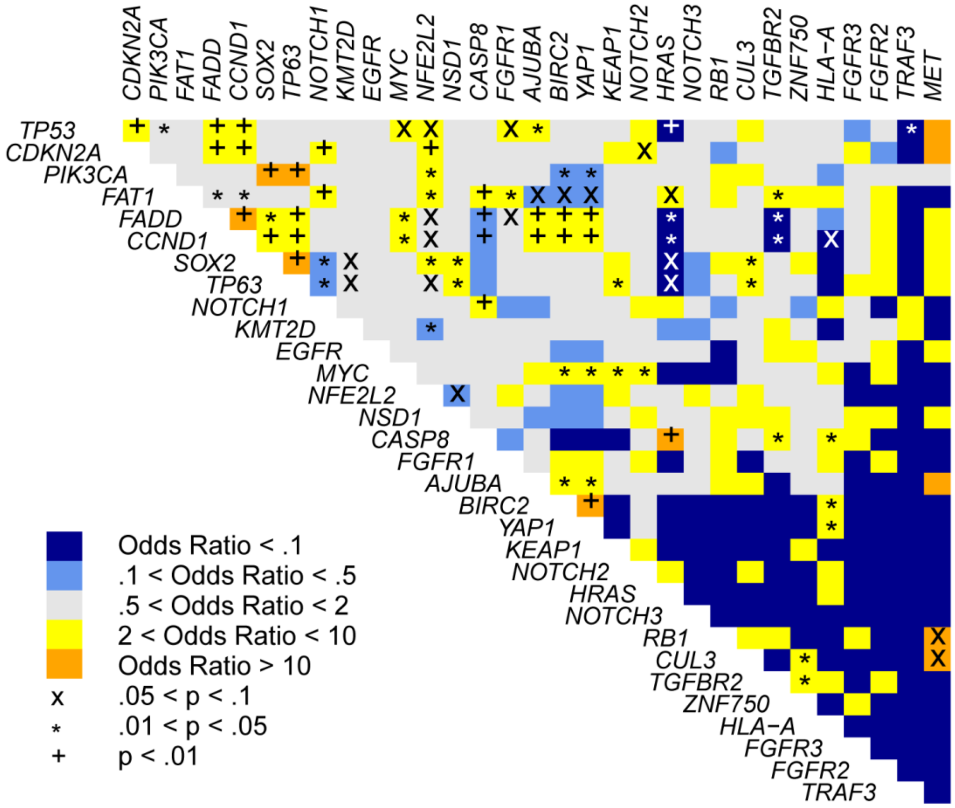


Figure S7.1. Co-occurrence and mutual exclusivity of select genomic events. Odds ratios for occurrence of genomic events involving pairs of genes are displayed in the heatmap and colored according to the value of the odds ratio. Mutually exclusive events produce odds ratios less than one, while co-occurring events produce odds ratios greater than one. The evidence for mutual exclusivity (co-occurrence) increases as the odds ratio decreases (increases). p-values were derived from a cBioPortal Mutual Exclusivity Modules analysis.

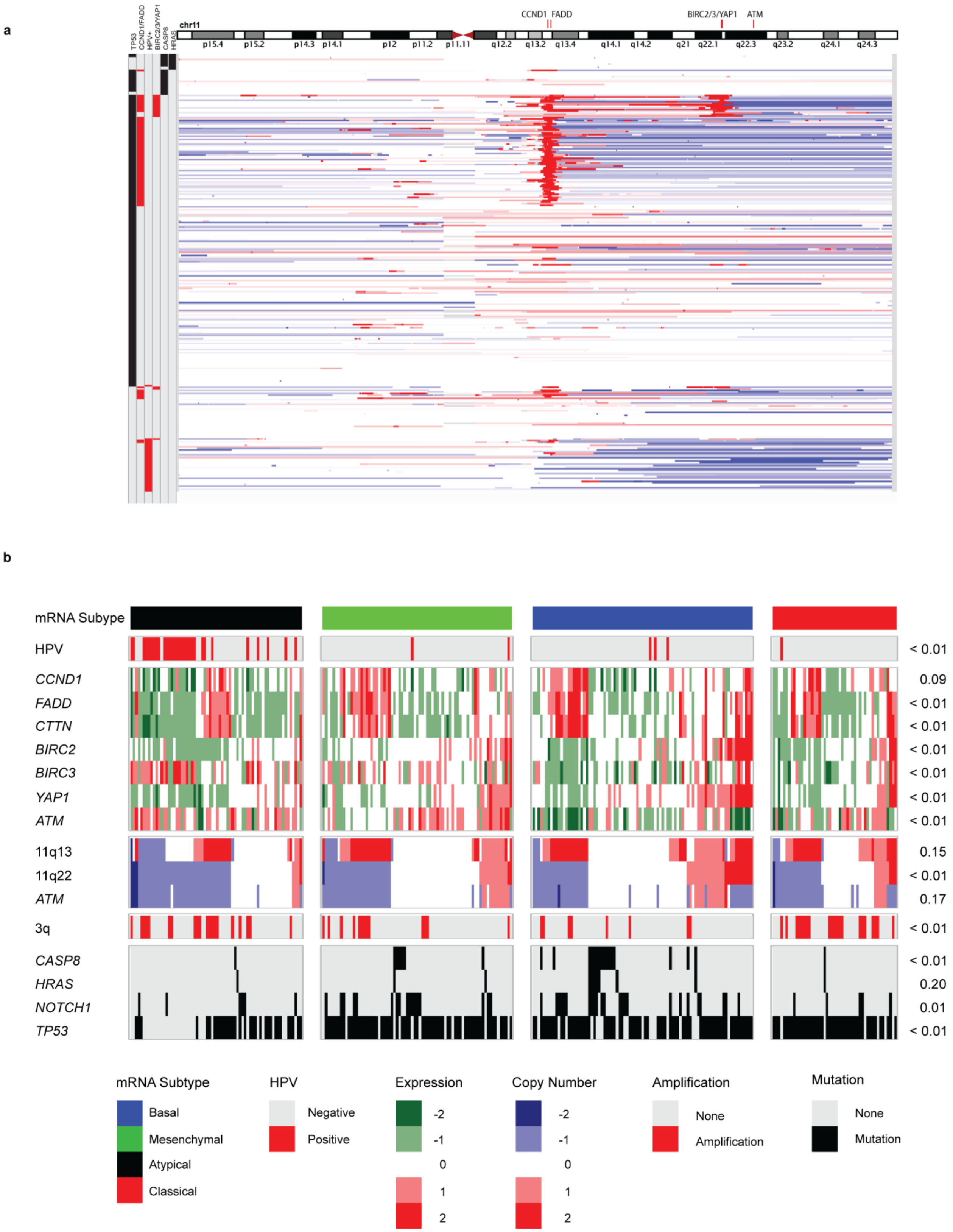


Figure S7.2. DNA copy number and gene expression in chromosome 11q. (a) Modified Integrated Genome Viewer plot for chromosome 11 is shown with a diagram of the chromosome and the locations of key genes at

the top. Red indicates increased copy number and blue indicates reduced copy number. Each sample is a row and key genomic alterations are shown on the left; mutations for *TP53*, *CASP8* and *HRAS*, high level copy number gains for *CCND1/FADD* and *BIRC2/3/YAP1*. (b) Heatmap display of genomic events in chromosome 11 documenting that focal “high-level” amplification and expression most commonly seen in the basal subtype, and in combination with deletion and low expression of genes in the telomeric regions of 11q22 including *ATM*. Additionally shown is the anti-correlation of 11q amplifications with *CASP8* and *HRAS*, primarily in the basal subtype. Samples are arranged in columns and organized by gene expression subtype; select variables appear in rows and assume values indicated in the heatmap legend. Discrete gene expression measurements were produced by standardizing the RNA-Seq gene quantification measurements within each gene and then binning. Discrete gene copy number values were obtained from the GISTIC output, as described in S2.1. 3q amplification is defined to be high level gain – i.e. discrete gene copy number equal to 2 – for at least one of *PIK3CA*, *SOX2*, or *TP63*. Fisher’s exact test p-values at the right of the figure show associations between the expression subtypes and the genomic events. Monte Carlo versions of Fisher’s exact test were used when examining associations between the expression subtypes and either the discrete gene expression or discrete copy number values.

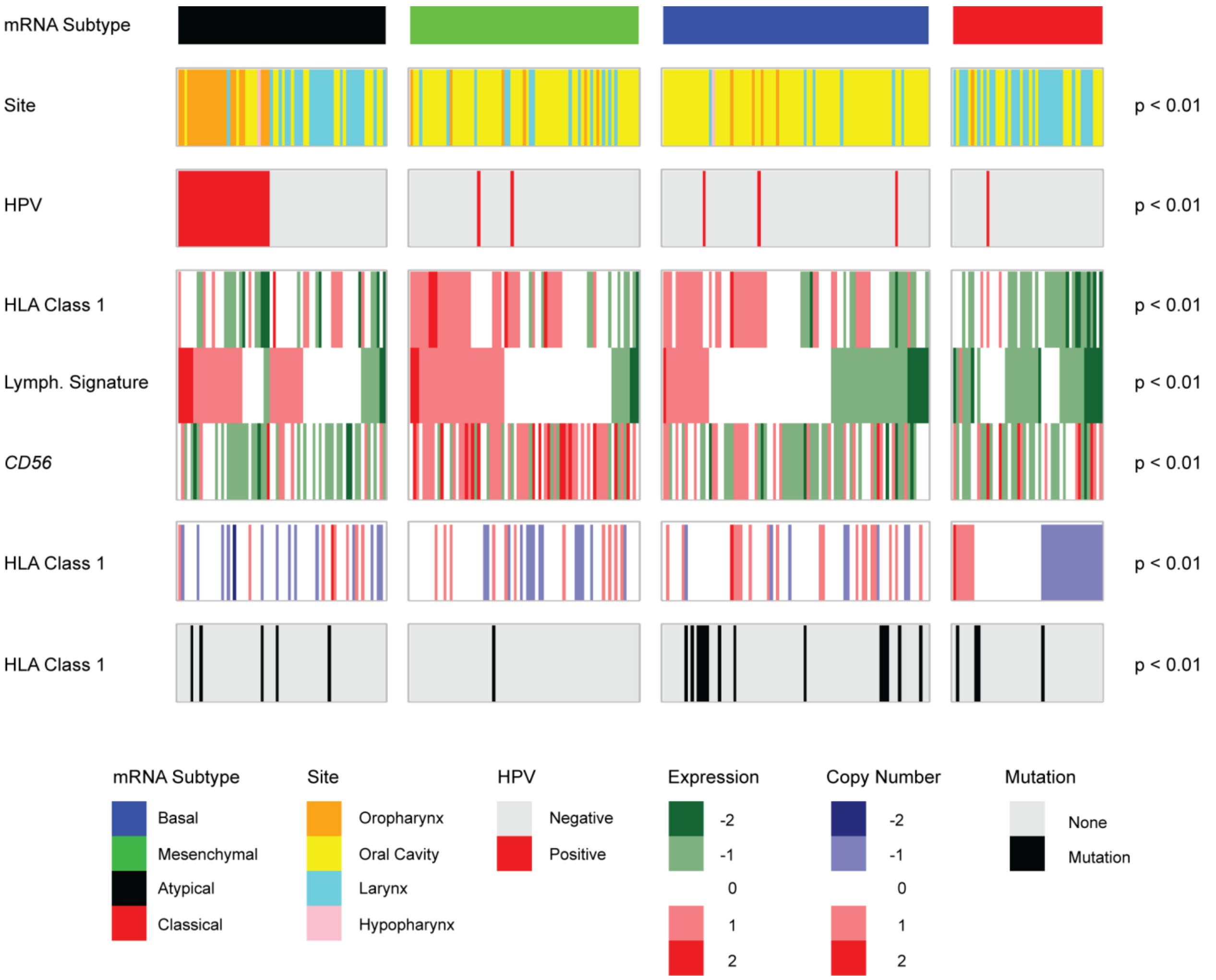


Figure S7.3. DNA copy number and gene expression for HLA class 1 and lymphocyte signature genes. Samples are arranged in columns and organized by gene expression subtype; select variables appear in rows and assume values indicated in the heatmap legend. Discrete gene expression measurements were produced by standardizing the RNA-Seq gene quantification measurements within each gene and then binning. Discrete gene copy number values were obtained from the GISTIC output, as described in Section S2.1. P-values measure the statistical significance of associations of the values in each row with mRNA subtype, and these were assessed with a Monte Carlo version of Fisher's exact test using 10,000 simulations.

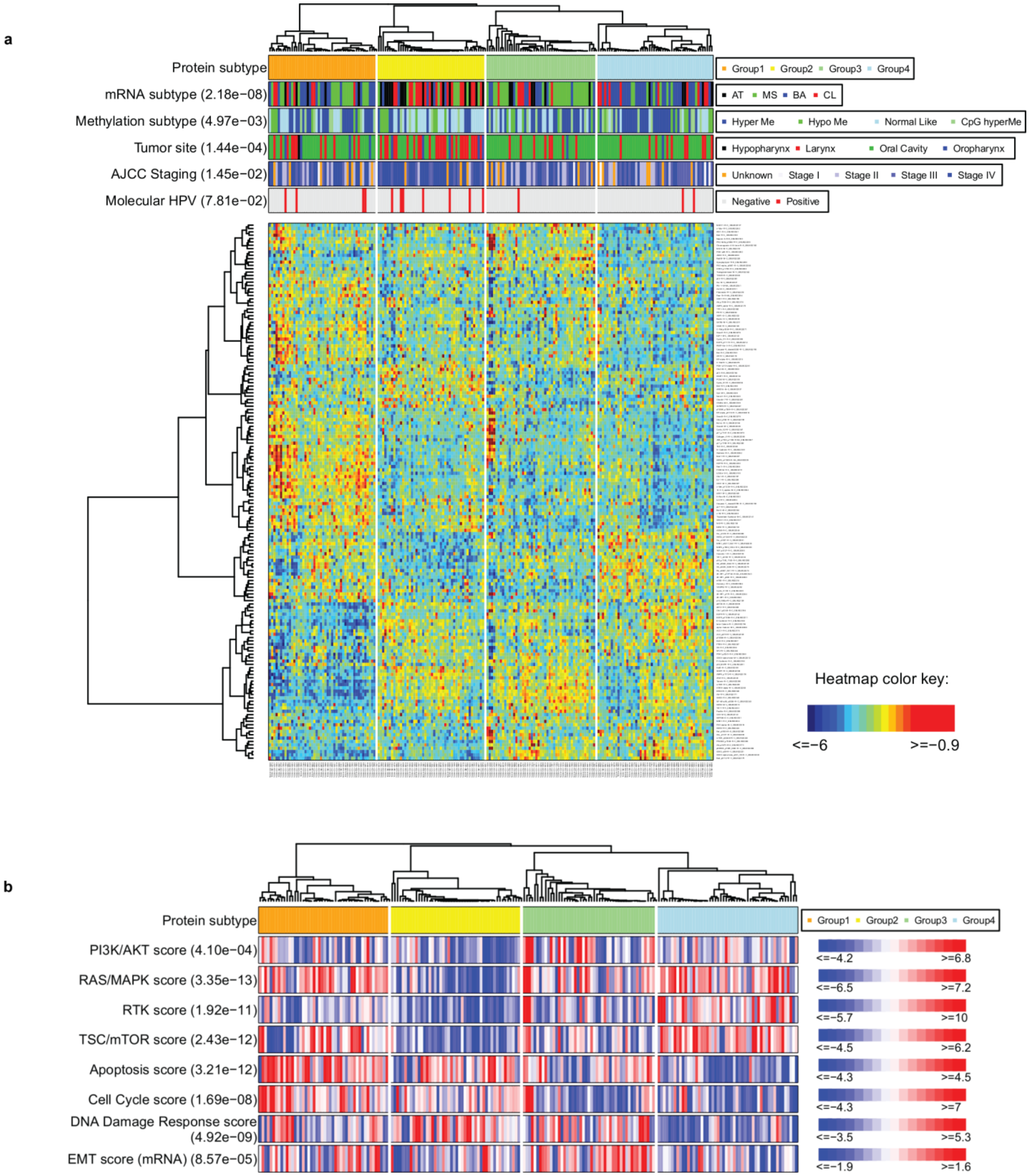


Figure S7.4. Unsupervised clustering of reverse phase protein array data by non-negative matrix factorization (NMF) clustering. (a) NMF clustering was applied to the reverse phase protein array data, identifying 4 subgroups of tumors with distinct protein expression patterns. RPPA subgroups were significantly associated with subtypes identified in independent data types (including mRNA clustering and methylation subtypes) and with clinical characteristics (tumor site, stage, and grade). (b) Clustering by RPPA correlates with pathway scores derived from protein data including EMT signature (bottom row). mRNA subtypes: AT, atypical; MS,

mesenchymal, BA, basal; CL, classical; methylation subtypes: Hyper Me, hypermethylated, Hypo Me, hypomethylated, CpG hyperme, CpG island hypermethylated.

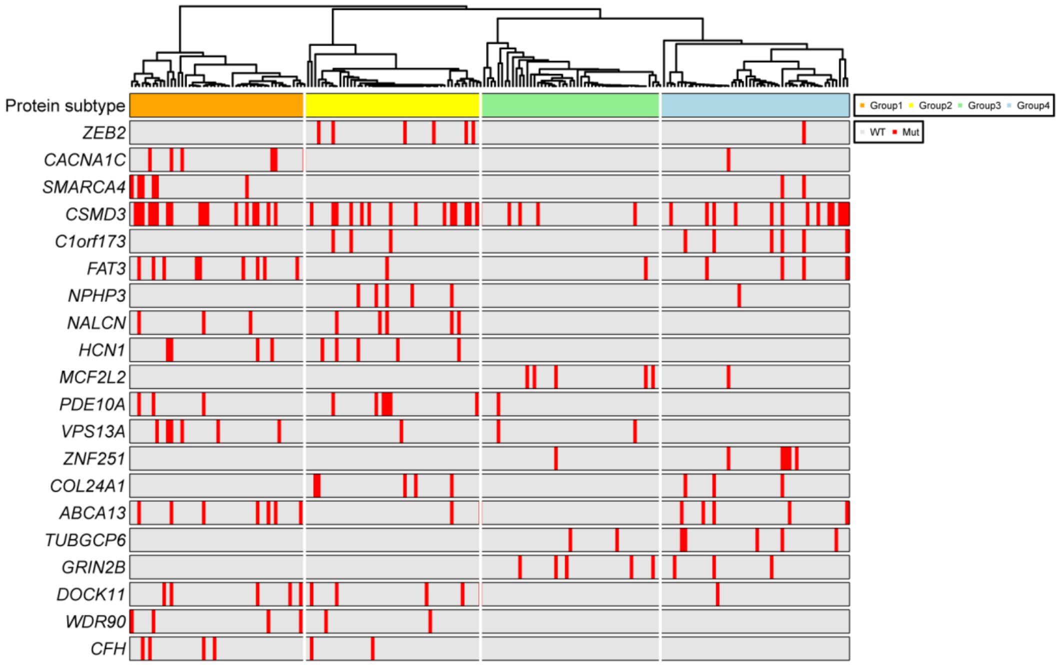


Figure S7.5. Correlation of RPPA subtypes (by NMF clustering) and mutations. Twenty genes with mutations with high correlation to RPPA subtypes are shown ($p < 0.05$). Statistical testing by ANOVA to compare differences in molecular and clinical data between the 4 proteomic subgroups identified by RPPA. WT, wildtype; Mut, mutation.

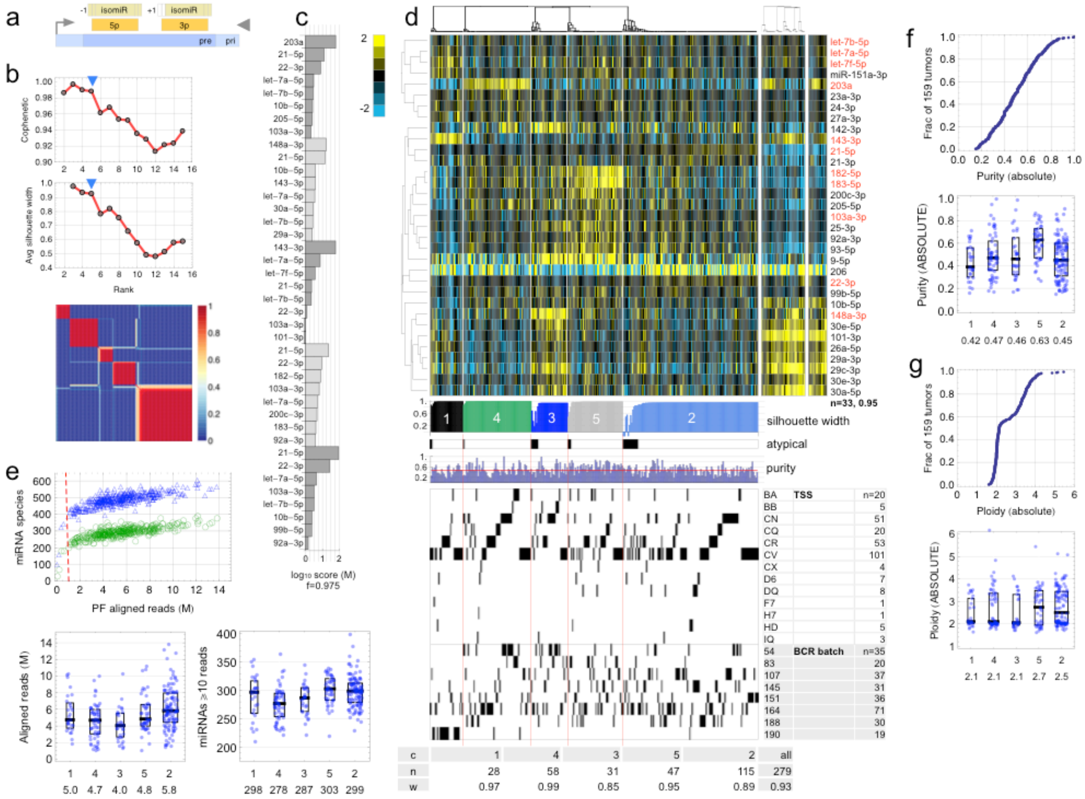


Figure S7.6. Unsupervised clustering of miRNA-Seq data. (a) Schematic of an miRNA primary transcript (pri), the trimmed pre-miRNA (pre), reference 5p and 3p strands, and potential 5' and 3' isomiR variation. The gray triangle points to the 5p/3p-strand data representation used. (b) From the NMF consensus clustering rank survey[100], both cophenetic correlation coefficient and average silhouette width suggest a 5 group solution. The blue/red consensus membership heatmap indicates that groups in this solution contain relatively few 'atypical' members (yellow-white). (c) Discriminatory miRs with the largest 2.5% of scores in each metagene, log10 scale. (d) NMF consensus clustering. Top to bottom: normalized abundance heatmap for the 33 5p or 3p strands that were most discriminatory for NMF (i.e. the top 5% of metagene scores), for 279 tumor samples, 37 adjacent normal mucosa samples, and 16 selected adjacent normal mucosa samples; silhouette width profile; "atypical" group members, which are samples with a width below 0.9 of the maximum in a group; sample purity (see f) and covariate tracks showing tissue source site and BCR batch number, with p-values were calculated using a Fisher exact test for TSS and a Chi-squared test for BCR batch; summary table of group number (c), number of samples (n) and average silhouette width (w). The scale bar shows log2 normalized abundances, which are RPMs values, median-centered for each miR in the heatmap. The 33 miRs shown correspond to the largest 5% of scores in each metagene. Red text highlights the more discriminatory subset. (e) Above: The number of sequencing reads aligned to miRBase annotations and the number of miRBase annotations with at least 1 (blue) or 10 (green) reads aligned. Below: Per-group distributions of the number of sequencing reads aligned to miRBase annotations, and the number of miRBase annotations with at least 10 aligned reads. (f,g) Distribution functions and per-group distributions for sample purity and ploidy as estimated by ABSOLUTE[29].

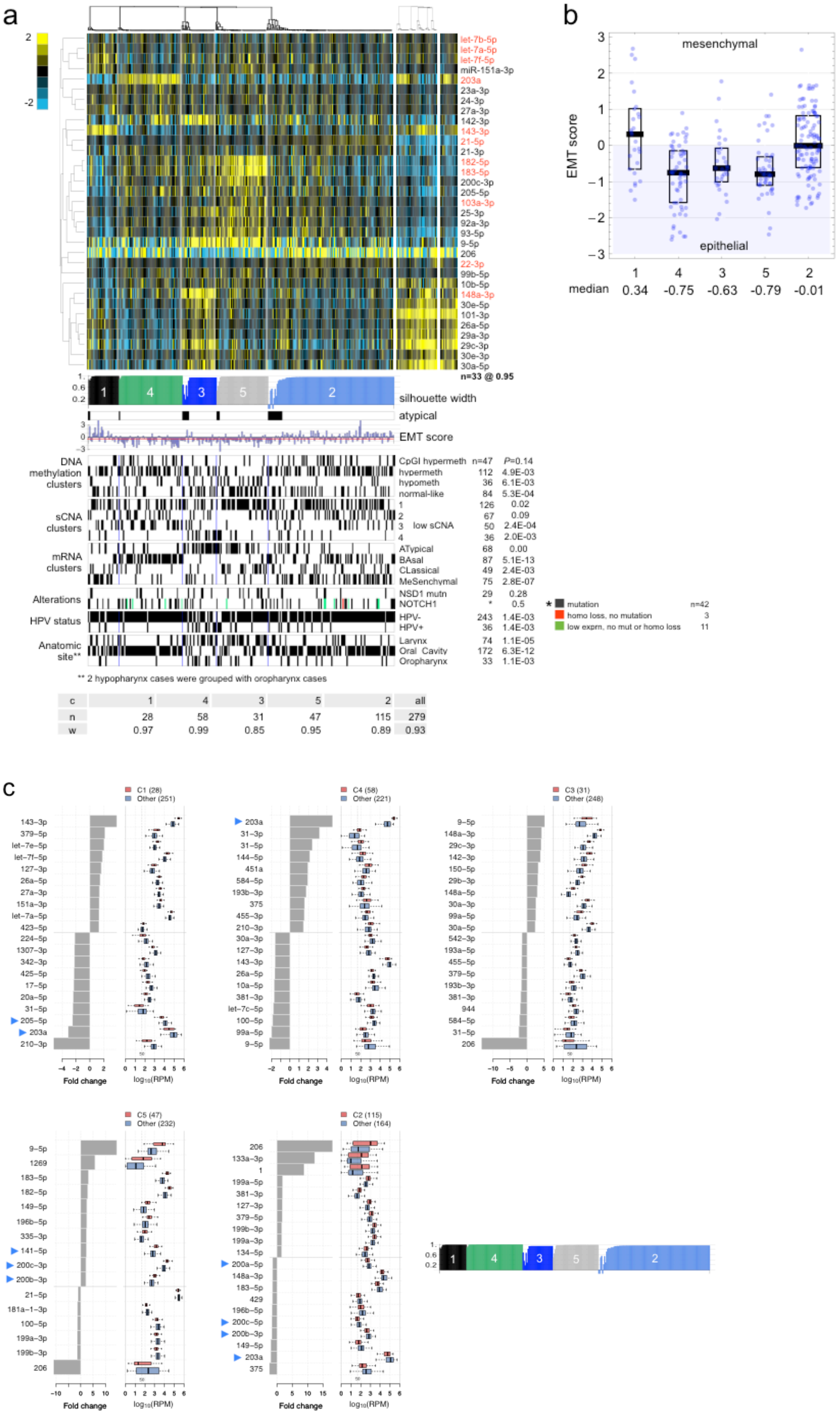


Figure S7.7. Covariates, EMT scores and differentially abundant miRNAs by unsupervised cluster. (a) A normalized abundance heatmap for NMF clusters, as in Figure S7.6. Below this is a profile of EMT scores[98], then covariate tracks showing DNA methylation clusters, somatic copy number clusters (with the low-SCNA cluster highlighted), mRNA-seq clusters, *NSD1* mutations and *NOTCH1* alterations, HPV status, and three

anatomic sites. P-values are from Fisher exact tests. (b) Distributions of EMT scores[98]. (c) miRNA 5p or 3p strands that are differentially abundant between samples in each tumor cluster vs. all other tumor samples. Left: fold change, linear scale. Right: distributions of RPM abundance, log scale. Up to 10 of the largest fold changes in each direction are shown; FDR ≤ 0.05 (see Methods). Blue triangles mark miRs that are known to be associated with EMT[158]. Homo loss, homozygous loss; low exprn, no mut or homo loss, low expression, no mutation or homozygous loss.

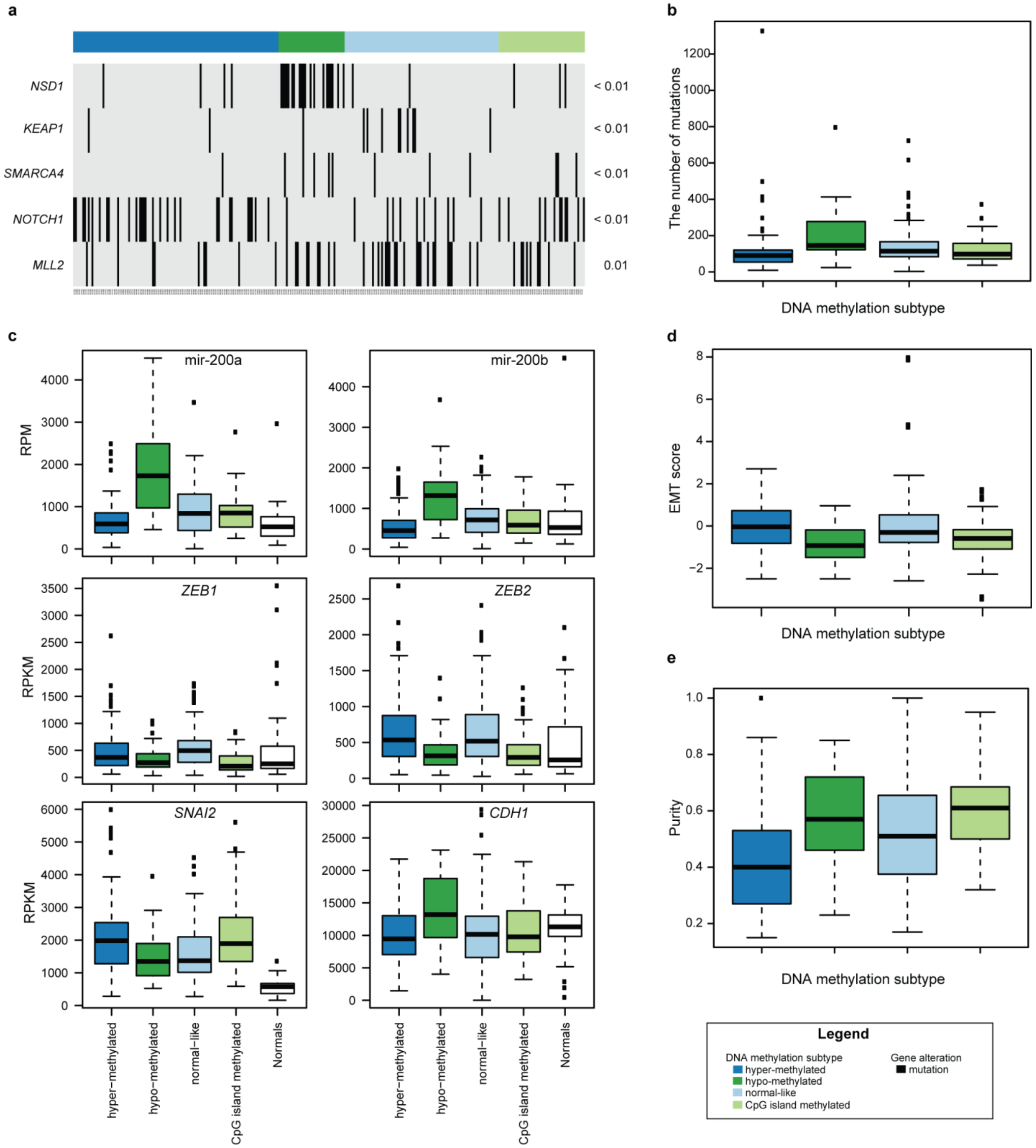


Figure S7.8. DNA methylation subtypes are associated with somatic mutations, EMT score, and target gene expression. (a) Mutations are shown in black. (b) Total number of non-synonymous mutations, colors correspond to top bar of heatmap on Figure 4b. (c) Standardized miRNA and mRNA read counts are depicted in boxplots, colors correspond to top bar of heatmap on Figure 4b. (d) EMT score is derived using RPPA markers.

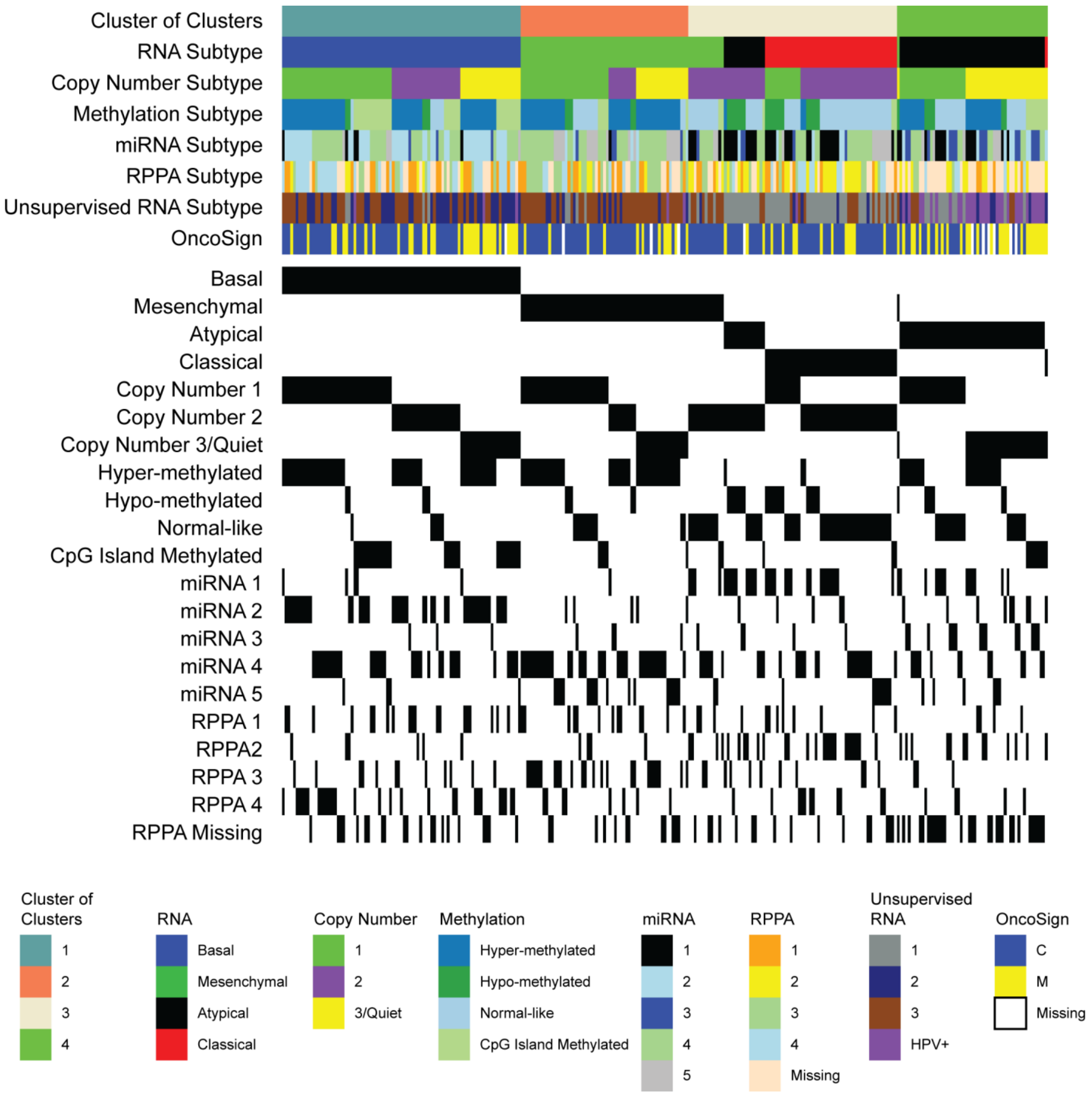


Figure S7.9. Cluster of clusters analysis. Class labels corresponding to RNA, DNA copy number, DNA methylation, miRNA, and RPPA subtypes were analyzed using unsupervised clustering techniques, as described in S7.3. The resulting four "Cluster of cluster" class labels are displayed along with annotation tracks corresponding to class labels for the various genomic platforms. These findings show the presence of concordant patterns detected in multiple genomic platforms, including underlying associations with tumor site and HPV status.

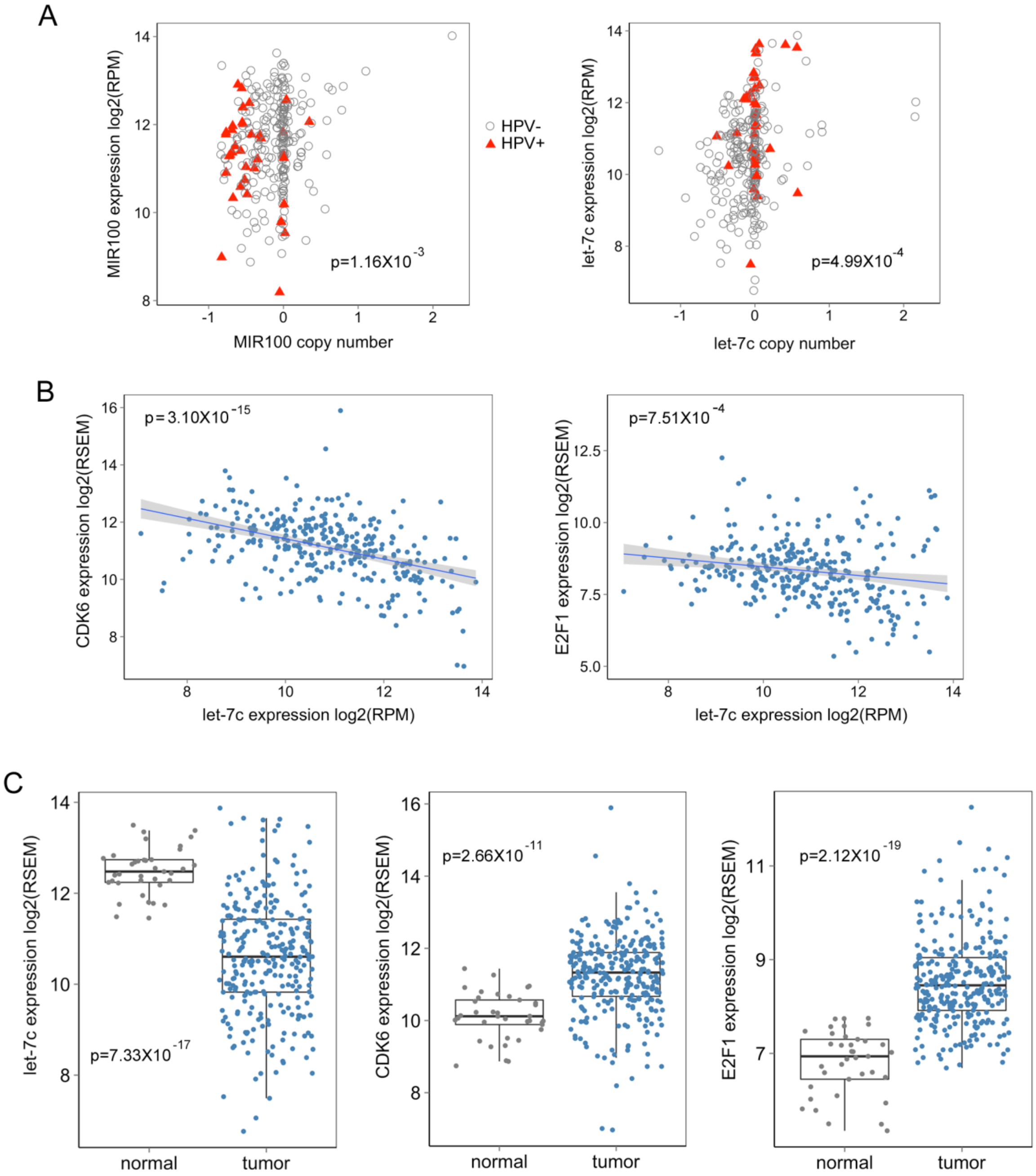


Figure S7.10. Decreased copy-number and expression of miR-100-5p and let-7c-5p are correlated with increased *CDK6* and *E2F1* expression in head and neck cancer. (A) miR-100-5p and let-7c-5p expression is correlated with DNA copy-number variation (CNV). The correlation of miR-100-5p and let-7c-5p expression with their DNA CNVs was analyzed in 279 tumor specimens. *P*-values give the significance of the linear regression of miRNA expression as a function of CNV. Tumor specimens include 243 HPV(-) (gray open circles), and 36 HPV(+) samples (red solid triangles). When we perform the analysis for HPV(+) and HPV(-) samples separately, association is only significant in HPV(-) samples. (B) A strong negative correlation between let-7c-5p expression and expression of its validated targets *CDK6* [159] and *E2F1* [160]. The shaded band

represents the 95% confidence interval on the best-fit line. *P*-values from linear regression are given. (C) Inverse relation between expression of miRNA let-7c-5p and its targets *CDK6* and *E2F1*. Expression of miRNA let-7c-5p in 279 tumor specimens (blue dots) was lower than in 37 adjacent mucosa tissues (gray dots, left panel). Expression of *CDK6* and *E2F1* in tumor specimens was higher than in adjacent mucosa tissues (middle and right panels). *P*-values from Mann-Whitney-Wilcoxon tests are given. Box-whisker plots show medians, inter-quartile ranges (boxes), and lines that extend to 1.5 times the inter-quartile range.

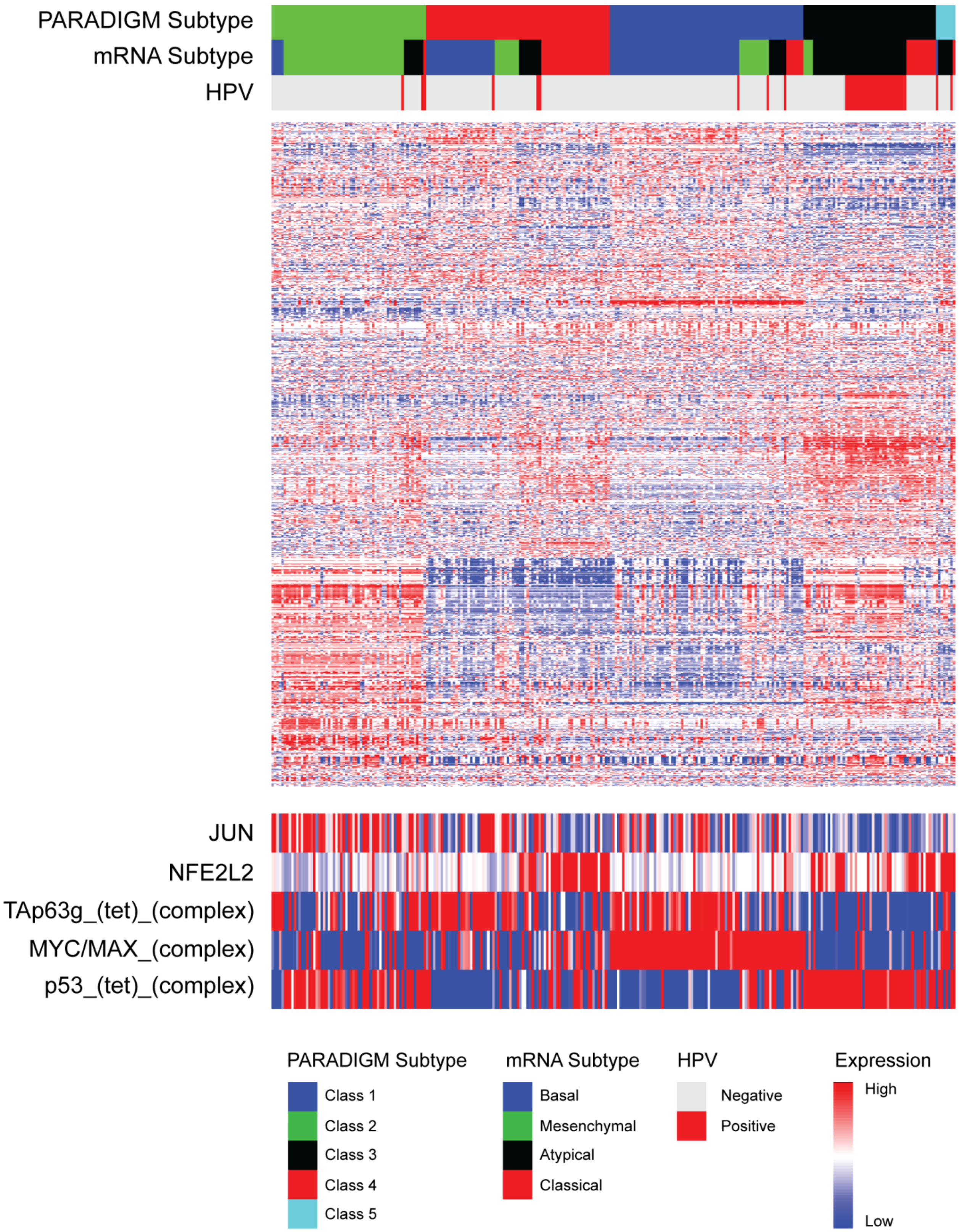


Figure S7.11. Subtypes defined by PARADIGM integrated pathway levels. Heatmap display of the top 25% varying IPLs after consensus clustering (top) and select integrated pathway levels (bottom). Samples are arranged in order of their consensus cluster, with expression subtype and molecular HPV status displayed in the column header.

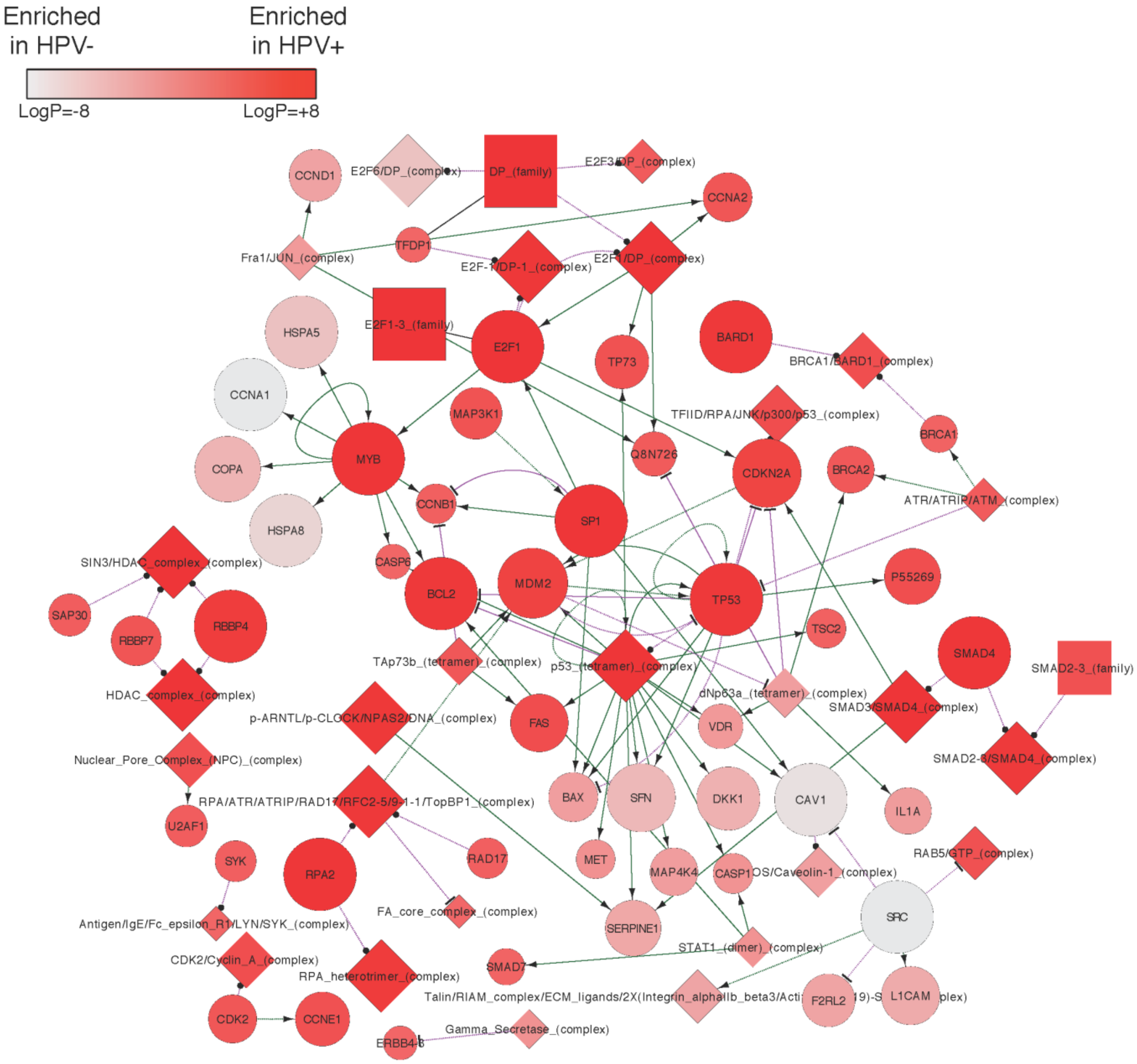
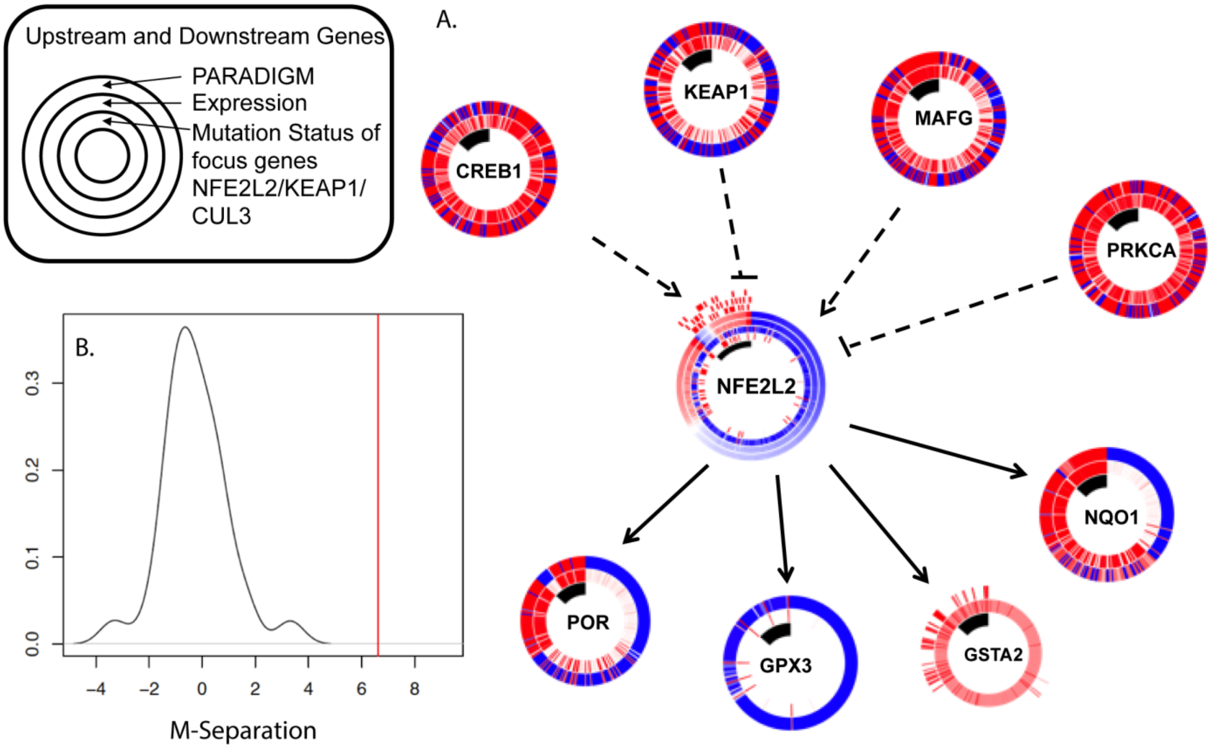
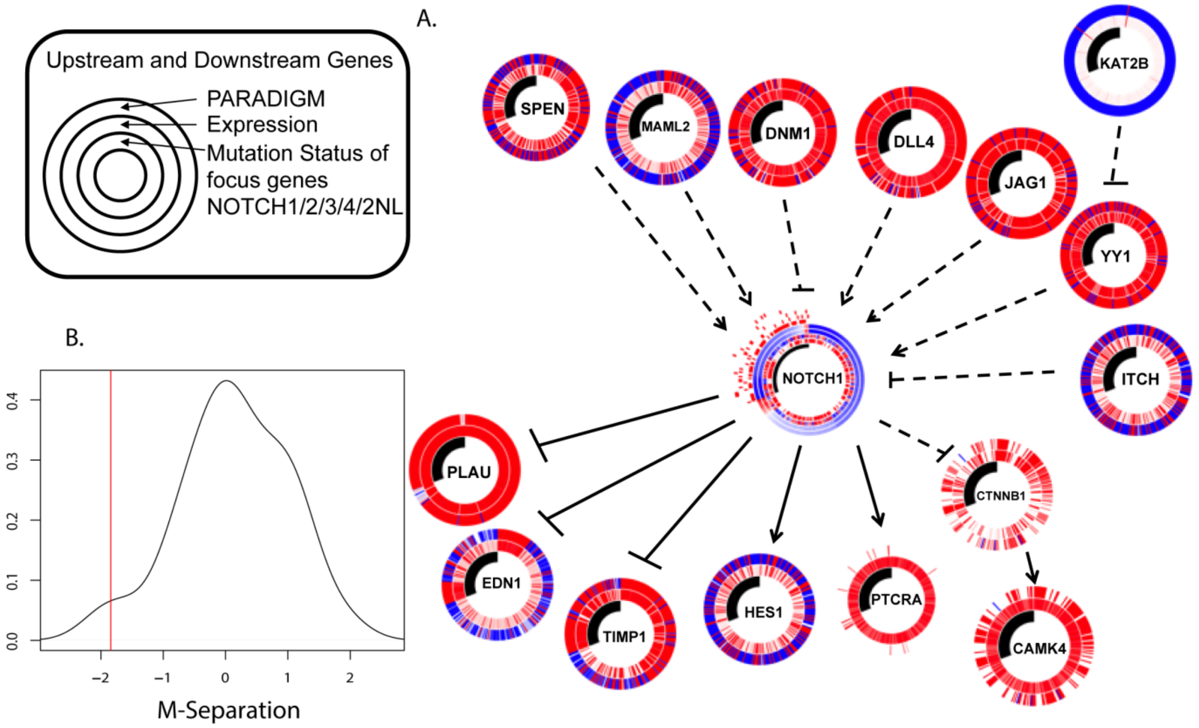


Figure S7.12. Enriched sub-network for features significantly differentiated between HPV(+) and HPV(-) samples. Red nodes are those features that had significantly higher means across the HPV(+) samples, and nodes in light gray were found to be significantly higher in HPV(-) samples. Solid links indicate transcription targets while kinase signaling is illustrated with dashed links. Green links are those annotated to be activating, while magenta links are annotated as inhibiting.



NFE2L2 Rings (Inner to Outer) : NFE2L2 Pathway Mutations, Expression, PARADIGM (Upstream), PARADIGM (Downstream), PARADIGM-shift (D-U), NFE2L2 Mutation, KEAP1 Mutation, CUL3 Mutation

Figure S7.13. PARADIGM-SHIFT analysis of *NFE2L2*. (A) Circle maps for pathway features of the local Paradigm-Shift network that infer gain of function *NFE2L2* pathway mutations. Each ring shows multiple types of data for the mutated versus non-mutated samples, indicated by the innermost black arc. (B) Distribution of the real shifts versus the null shifts, showing significant shift in the gain-of-function direction.



NOTCH Rings (Inner to Outer): NOTCH Family Mutations, Expression, PARADIGM (Upstream), PARADIGM (Downstream), PARADIGM-shift (D-U), NOTCH1 Mutation, NOTCH2 Mutation, NOTCH3 Mutation, NOTCH4 Mutation, NOTCH2NL Mutation

Figure S7.14. PARADIGM-SHIFT analysis of NOTCH family genes. (A) Circle maps for pathway features of the local Paradigm-Shift network that infer loss of function NOTCH family mutations. Each ring shows multiple types of data for the mutated versus non-mutated samples, indicated by the innermost black arc. (B) Distribution of the real shifts versus the null shifts, showing significant shift in the loss-of-function direction.

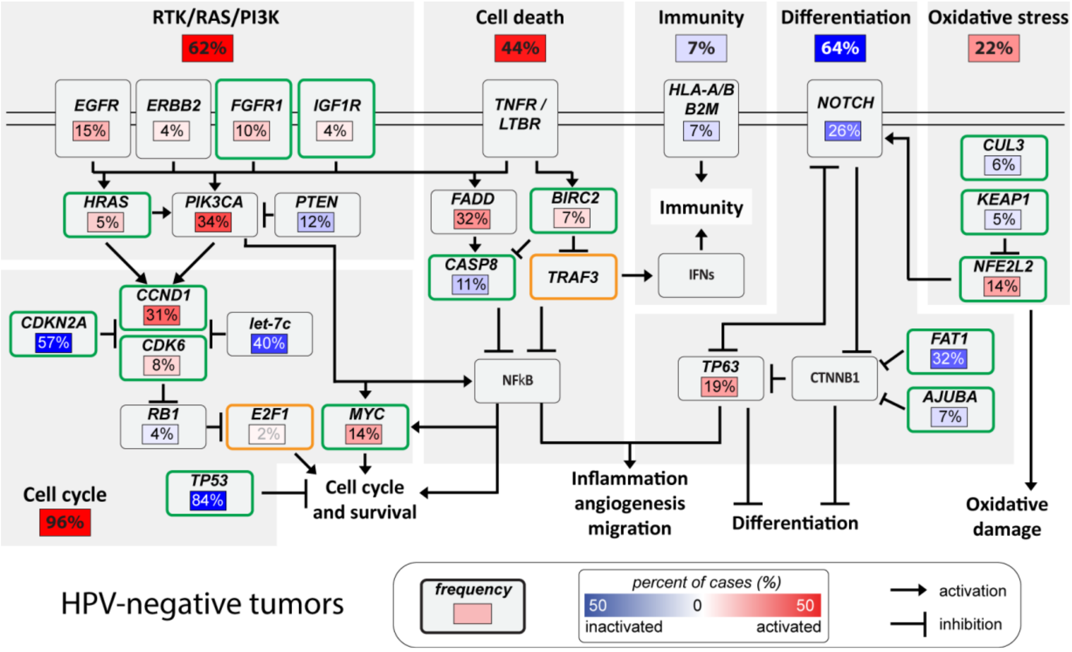


Figure S7.15, part 1. Diversity and frequency of genetic changes leading to deregulation of signaling pathways and transcription factors in HPV (-) tumors.

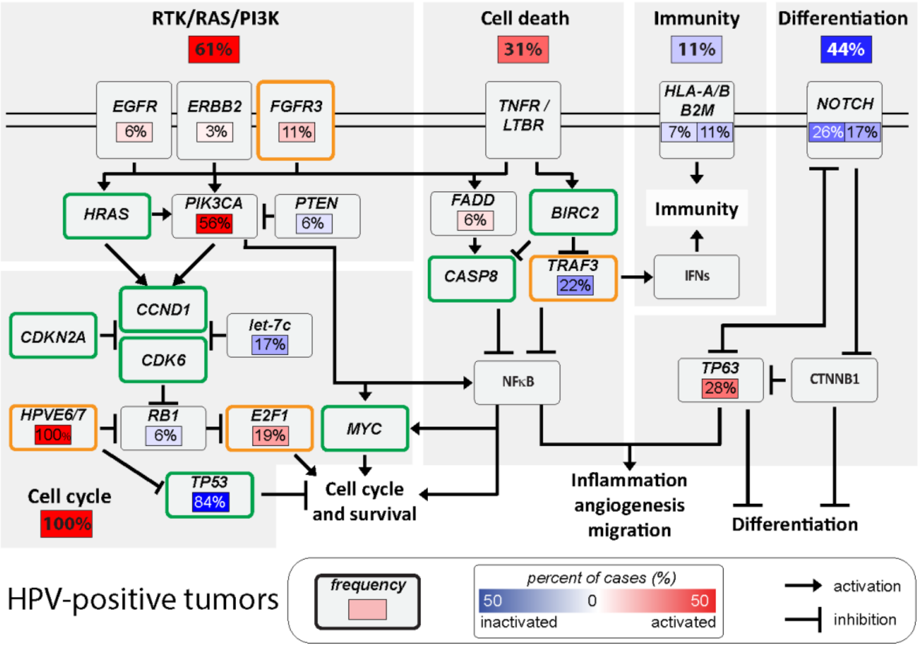


Figure S7.15, part 2. Diversity and frequency of genetic changes leading to deregulation of signaling pathways and transcription factors in HPV(+) tumors.

Figure S7.15. Diversity and frequency of genetic changes leading to deregulation of signaling pathways and transcription factors in HPV (-), part 1, and HPV(+),part 2 HNSCC. The frequency (%) of genetic alterations for part 1, HPV(-) and part 2, HPV(+) tumors are shown. Key affected pathways (gray panels), components (box panels), and inferred functions, are summarized in the key and main text with supporting citations. Component alterations are classified under the pathways and include homozygous deletions, focal amplifications, and somatic mutations as in Data File S7.3. Those predominant in HPV(-) (green) or HPV(+)(orange) subsets are highlighted. Activated pathways / genes (red), inactivated pathways/genes (blue), activating arrows, or inhibitory symbols shown are inferred based on predicted effects of genome alterations and/or pathway functions as cited in the text. Pathway components with interactions coordinating cell cycle (*CCND1, CDK6*) or oxidative stress (*KEAP1, CUL3*) are shown without separation.

miRNA	t	P.Value	adj.P.Val
hsa-mir-200a	10.07031	1.57E-20	1.64E-17
hsa-mir-200b	7.715896	2.18E-13	1.14E-10
hsa-mir-429	6.997526	1.96E-11	6.83E-09
hsa-mir-3150b	6.590934	2.20E-10	5.74E-08
hsa-mir-504	6.505555	3.60E-10	6.39E-08
hsa-mir-675	6.502621	3.66E-10	6.39E-08
hsa-mir-1228	5.797659	1.83E-08	2.73E-06
hsa-mir-3166	5.627741	4.47E-08	5.20E-06
hsa-mir-219-1	5.627548	4.47E-08	5.20E-06
hsa-mir-1296	5.492367	8.97E-08	9.39E-06
hsa-mir-141	5.29299	2.45E-07	2.32E-05
hsa-mir-1227	5.275998	2.67E-07	2.32E-05
hsa-mir-551a	5.238532	3.21E-07	2.58E-05
hsa-mir-183	5.154669	4.84E-07	3.61E-05
hsa-mir-1910	4.849994	2.06E-06	0.0001437
hsa-mir-549	4.679941	4.49E-06	0.0002937
hsa-mir-34c	4.640273	5.37E-06	0.0003095
hsa-mir-577	4.636511	5.46E-06	0.0003095
hsa-mir-940	4.630115	5.62E-06	0.0003095
hsa-mir-188	4.581508	6.98E-06	0.0003653
hsa-mir-149	4.40964	1.48E-05	0.0007385
hsa-mir-9-3	4.326441	2.12E-05	0.0009775
hsa-mir-182	4.322883	2.15E-05	0.0009775
hsa-mir-934	4.290453	2.47E-05	0.0010747
hsa-mir-2277	4.266262	2.73E-05	0.0011274
hsa-mir-331	4.254748	2.87E-05	0.0011274
hsa-mir-1266	4.251063	2.91E-05	0.0011274
hsa-mir-200c	4.236849	3.09E-05	0.0011538
hsa-mir-1180	4.179933	3.91E-05	0.0014115
hsa-mir-1226	4.145005	4.52E-05	0.0015757
hsa-mir-23a	4.094486	5.56E-05	0.0018747
hsa-mir-760	4.05732	6.46E-05	0.0020653
hsa-mir-96	4.055172	6.52E-05	0.0020653
hsa-mir-708	4.041552	6.88E-05	0.0021177

Table S7.1. miRNAs associated with *NSD1*-depleted/hypomethylated cluster

miRNA	mRNA	2FC(%)	Inverse	Del	Amp	Diploid	p-value
miR-100	*PLK1*	56	Y	83	57	15	5.59x10^{-3}
			N	41	63	16	
let-7c	*CDK6*	52	Y	68	13	62	9.11x10^{-4}
			N	34	17	81	
	E2F1	57	Y	72	15	79	2.30x10^{-3}
			N	30	15	73	
	HMGA2	71	Y	86	18	91	4.98x10^{-4}
			N	16	12	52	
	IGFBP1/2	72	Y	86	19	94	4.39x10^{-4}
			N	16	11	49	

Table S7.2. Increased mRNA expression associated with decreased miR-100 and let-7c expression in deleted genomic regions. 2FC = 2-fold differences. 2FC (%) indicates the percentage of tumor samples that exhibit an inverse relationship between a miR and its target gene, wherein both the miR and mRNA exhibit at least a 2--fold change in expression between tumor and adjacent mucosa samples. “Inverse” denotes cases of tumor specimens that display a 2-fold change inverse relationship (Y) in the deleted (Del), amplified (Amp), or diploid regions. “N” represents those cases that don’t satisfy a 2-fold change inverse relationship. A Fisher exact test was performed to determine the significance of association between miRNA copy number loss and fold change for the miR-mRNA pair.

HPV status	Gene	Cytoband	Deletions (%)		q-value
			Homozygous	Heterozygous	
HPV(-)	mir-100	11q23.1	0	101 (41.6%)	2.31×10^{-9}
	let-7c	21q21.1	2 (0.8%)	94 (38.7%)	0.15
HPV(+)	mir-100	11q23.1	2 (5.6%)	25 (69.4%)	1.93×10^{-7}
	let-7c	21q21.1	0	6 (16.7%)	>0.25

Table S7.3. Copy number loss of miR-100 and let-7c in tumor specimens. The analysis of copy number variation was performed in a total of 279 tumor specimens, including 243 HPV negative, and 36 HPV positive tumors. For this analysis q-value represents false discovery rates for the aberrant regions (q-values below 0.25 are considered significant).

S8. DNA methylation profiling

Sample preparation and hybridization

The Illumina Infinium HumanMethylation450 [161] array was used for TCGA HNSCC samples. This platform includes probes for more than 480,000 CpG sites, spanning 99% of RefSeq. In total, 96% of CpG islands and 92% of CpG shores are represented by at least one probe. This array is an expansion of the Illumina Infinium HumanMethylation27 array [162] previously used in TCGA, which interrogates 27,578 CpG dinucleotides spanning 14,495 unique gene regions, heavily concentrated near CpG islands.

Genomic DNA (1000 ng) for each sample was treated with sodium bisulfite, recovered using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA) according to the manufacturer's specifications and eluted in 18 µl volume. An aliquot (3 µl) is removed for MethyLight-based quality control testing of bisulfite conversion completeness and the amount of bisulfite converted DNA available for the Infinium DNA Methylation assay as described in Davis et al. [163]. All TCGA DNA samples passed quality control and proceeded to the Infinium DNA methylation assay. Each bisulfite-converted DNA sample was whole genome amplified (WGA) followed by enzymatic fragmentation as specified by the manufacturer. The bisulfite-converted, fragmented WGA-DNA samples were then hybridized overnight to a 12 sample BeadChip. During this hybridization, the WGA-DNA molecules anneal to methylation-specific DNA oligomers linked to individual bead types, with each bead type corresponding to a specific DNA CpG site and methylation state. The oligomer probe designs follow the Infinium I and II chemistries, in which locus-specific base extension follows hybridization to a methylation-specific oligomer. There are two different bead types for each locus, one with an oligomer that anneals specifically to the methylated version of the locus, while the other oligomer anneals to the unmethylated version of the locus. The Infinium I probes terminate complementary to the interrogated CpG site for methylated loci, or complementary to the TpG for unmethylated alleles. A matched oligomer-template DNA molecule hybrid will allow for the incorporation of a labeled nucleotide immediately adjacent to the interrogated CpG (or TpG) site. However, if the probe and template are mismatched, then primer extension will not occur. Adenine and thymine nucleotides are labeled with cy5 (red), while cytosine nucleotides are labeled with cy3 (green). No insertion of guanine nucleotides occurs in Inifnium I assays. Of note, the identity of the dye is representative of the nucleotide adjacent to the CpG dinucleotide. The methylation discrimination is derived from separate measurements from the two different types of beads present for each locus. For some loci, both measurements will be cy3, and for others both will be cy5. The Infinium type II chemistry is a true two-color system. A matched oligomer-template DNA molecule hybrid will allow for the incorporation of a labeled nucleotide immediately 3' to the interrogated CpG (or TpG) site. Adenine nucleotides labeled with cy5 (red) are incorporated at unmethylated (TpG) sites, while guanine nucleotides labeled with cy3 (green) are incorporated at methylated (CpG) sites. The intensities of both cy3 and cy5 are obtained after scanning each hybridized array. BeadArrays are scanned and the raw data are imported into custom programs in R computing language for pre-processing and calculation of beta value DNA methylation scores for each probe and sample.

Data processing

Raw image files were imported into R[8] for pre-processing and calculation of beta value DNA methylation scores, using the methylumi Bioconductor package [163]. Pre-processing steps include background correction, dye-bias normalization, and calculation of beta values and detection p-values.

p16 methylation status

As described in a prior report of lung cancer using similar array platforms [31], we defined a binary methylation status call for the protein coding transcript of *CDKN2A*, p16. Briefly, a single probe was selected as the optimal representation of promoter methylation status of the exon commonly referred to as E1alpha (probe cg13601799). This probe is annotated to a genome position located between exon 1b and exon 1a. Based on prior work a beta value for this probe of greater than 0.15 is highly associated with absent expression of the E1alpha exon.

S9: miRNA sequencing

Library construction and sequencing

microRNA sequence (miRNA-seq) data were generated for 279 tumor and 37 adjacent mucosa specimens using methods previously described [99]. Briefly, two micrograms of total RNA per sample are arrayed into 96-well plates, with controls as described below. RNA entering library construction is required to have at least a minimum quality on the BCR submission documentation. Total RNA is mixed with oligo(dT) MicroBeads and loaded into a 96-well MACS column, which is then placed on a MultiMACS separator (Miltenyi Biotec, Germany); the separator's strong magnetic field allows beads to be captured during washes. From the flow-through, small RNAs, including miRNAs, are recovered by ethanol precipitation. Flow-through RNA quality is checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip.

miRNA-Seq libraries are constructed using a plate-based protocol developed at the British Columbia Genome Sciences Centre (BCGSC). Negative controls are added at three stages: elution buffer is added to one well when the total RNA is loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR brew mix to a final well just before PCR. A 3' adapter is ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation of 1 hour at 22^{o} C. This adapter is adenylated, single-strand DNA with the sequence 5' /5rApp/ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and is incubated at 37^{o}C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation is completed, first strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'- CAAGCAGAAGACGGCATACGAGAT-3'). This is the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix is made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and DMSO. The mix is distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) is used to transfer the PCR template (1st strand cDNA) and indexed 5' PCR primers into the brew mix plate. Each indexed 5' PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3', contains a unique six-nucleotide 'index' (shown here as Ns), and is added to each well of the 96-well PCR brew plate. PCR is run at 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a 5 min incubation at 72^{o}C. Quality is then checked across the whole plate using a Caliper LabChipGX DNA chip. PCR products are pooled, then are size selected to remove larger cDNA fragments and smaller adapter contaminants, using a 96-channel automated size selection robot that was developed at the BCGSC. After size selection, each pool is ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters are generated, and lanes are sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

Preprocessing, alignment and annotation

Briefly, the sequence data are separated into individual samples based on the index read sequences, and the reads undergo an initial quality control (QC) assessment. Adapter sequence is then trimmed off, and the trimmed reads for each sample are aligned to the NCBI GRCh37-lite reference genome. Below we describe these steps in more detail.

Routine QC assesses a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error is estimated by a method originally developed for SAGE.

Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, typically they are ~21 bp long, and so are shorter than the 31- bp read length. Given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the reference genome, 3' adapter sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5') 15 bases of the 3' adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.

The algorithm first determines whether a read sequence should be discarded as an adapter dimer by checking whether the 3' adapter sequence occurs at the start of the read. For reads passing this stage, the algorithm then tries to identify an exact 15-bp match anywhere within the read sequence. If it cannot, it then retries, starting from the 3' end, and allowing up to 2 mismatches. If the full 15bp is not found, decreasing lengths of adapter are checked, down to the first 8 bases, allowing one mismatch. If a match is still not found, from 7 bases down to 1 base is checked, with an exact match required. Finally, the algorithm will trim 1 base off the 3' end of a read if it happens to match the first base of the adapter. This is based on two considerations. First, it is preferable to get a perfect alignment than an alignment that has a potential one-base mismatch. Second, if only 1 base of adapter was found in the read sequence, the read is likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample's overall miRNA profiling result.

After each read has been processed, a summary report is generated containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v16 is 15 bp, any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome. BWA9 alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches are used. Based on comparing expression profiles of test libraries (data not shown), reads that fail the Illumina base-calling chastity filter are retained, while reads that have soft-clipped CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases (Table S9.1), and requiring a minimum 3-bp overlap between the alignment and an annotation. In annotating reads we address two potential issues. First, a single read alignment can overlap feature annotations of different types; second, a read can have up to three alignment locations, and each alignment location can overlap a different type of feature annotation. By considering heuristically determined priorities

(Table S9.1), we resolve the first issue by giving each alignment a single annotation. We resolve the second by collapsing multiple annotations to a single annotation, as follows.

If a read has more than one alignment location, and the annotations for these are different, we use the priorities from Table S9.1 to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, the read is flagged as cross-mapped, and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa- mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

From the profiling results for a tumor type, for a minimum of approximately 100 samples, we identify the depth of sequencing required to detect the miRNAs that are expressed in a sample by considering a graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs. For the current work, a library from a sequenced pool was required to have at least 750,000 reads mapped to miRBase annotations. For any sequencing run that fails to meet this threshold, we sequence the sample again to achieve at least the minimum number of miRNA-aligned reads.

Finally, for each sample, the reads that correspond to particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Quantification files include information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

Priority	Annotation type	Database
1	mature strand	miRBase v16
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	"unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	
6	snoRNA	UCSC small RNAs, RepeatMasker
7	tRNA	
8	rRNA	
9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	

13	coding exons with zero annotated CDS region length	UCSC knownGenes
14	3' UTR	
15	5' UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker Low complexity	
24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

Table S9.1. Priorities for resolving annotation ambiguities for aligned miRNA-Seq reads.

S10: Batch effects analysis

S10.1 Methods

Hierarchical clustering and Principal Components Analysis (PCA) were used to assess batch effects in the TCGA HNSCC data sets. Four data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Illumina Infinium HumanMethylation 450 microarray), mRNA sequencing (Illumina HiSeq), and SNP (Affymetrix GenomeWide SNP 6.0). All of the data were at TCGA level 3 as per prior convention for batch effects analysis [31]. The analysis was with respect to two variables; batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects

For hierarchical clustering, the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure was used. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. This procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for the four data sets follow.

S10.2 Results by platform

miRNA (RNA-Seq Illumina HiSeq)

Figures S10.1 – S10.3 show clustering and PCA plots for miRNA-Seq data. miRNAs with zero values were removed and the read counts were log2-transformed before generating the figures. The figures show a small batch effect by batch number 215. However, the magnitude of batch effects was small and batch correction was not required. Given the size of the effect the concern was that batch effects correction algorithms could mask important biological variation in the data, along with the technical variation.

DNA Methylation (Infinium HM450 microarray)

Figures S10.4 – S10.6 show clustering and PCA plots for the Infinium DNA methylation platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

RNASeqV2 (RNA-Seq Illumina HiSeq)

Figures S10.7 – S10.9 show clustering and PCA plots for the RNA-Seq platform. Genes with zero values were removed and the values were log2-transformed before generating the figures. There were small batch effects by one tissue source site, Johns Hopkins, but with the small number of samples outlier biology could not be excluded. Once again, the overall effect was not considered large enough to warrant batch effects correction for the type of analyses done in this paper.

SNP (Genome Wide SNP6 hg19)

Figures S10.10 – S10.12 show clustering and PCA plots for the Genome Wide SNP 6.0 platform. Segment values were mapped to gene values for the analysis using Hg19. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

S10.3. Conclusions

Batch effects were analyzed in four different data sets. miRNA and mRNA data showed small batch effects in some of the samples. However, the batch effects weren't considered strong enough to warrant algorithmic batch effects correction; DNA methylation or SNP 6.0 data didn't show any major batch effects.

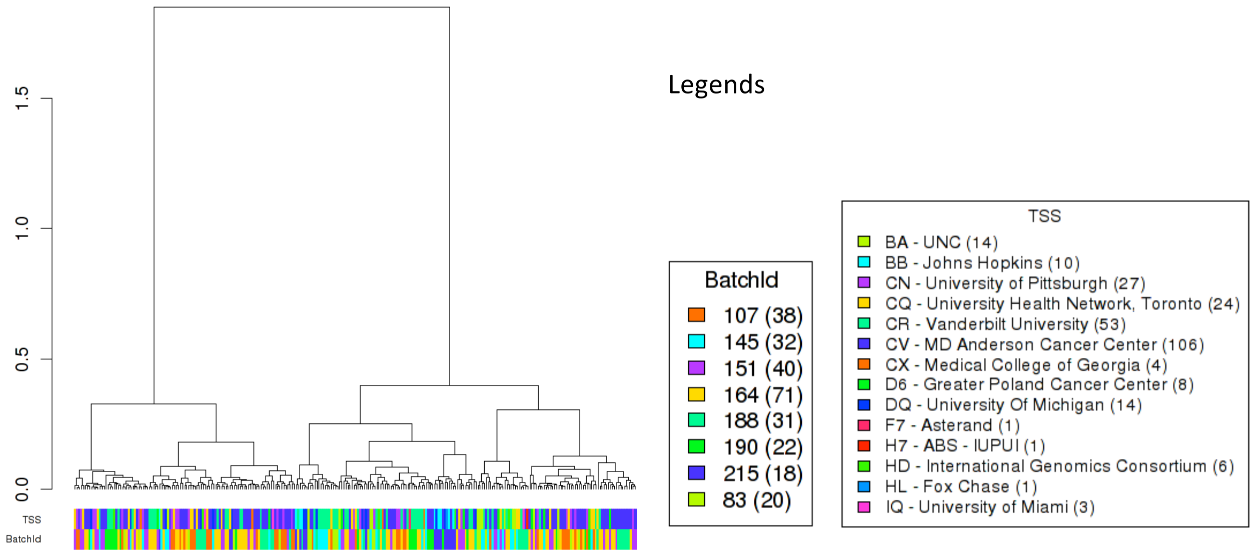
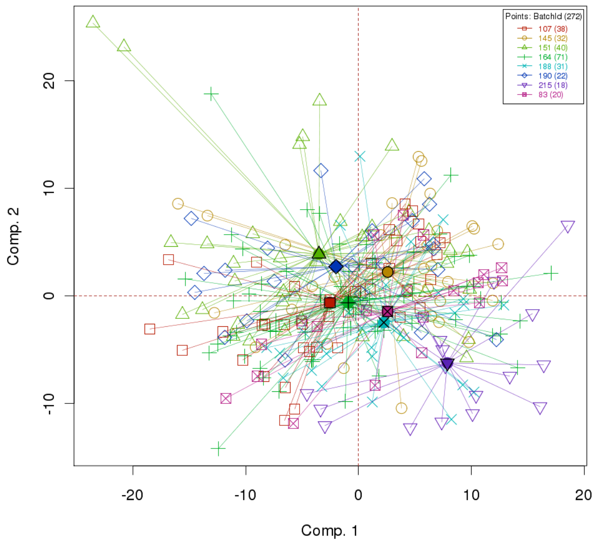


Figure S10.1 Hierarchical clustering for miRNA expression from miRNA-Seq data.



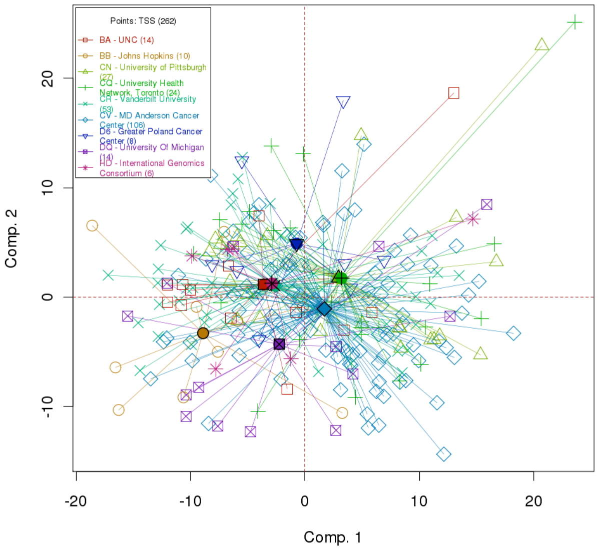


Figure S10.2 PCA: First two principal components for miRNA expression from miRNA-Seq data, with samples connected by centroids according to batch ID.

Figure S10.3 PCA: First two principal components for miRNA expression from miRNA-Seq data, with samples connected by centroids according to tissue source site.

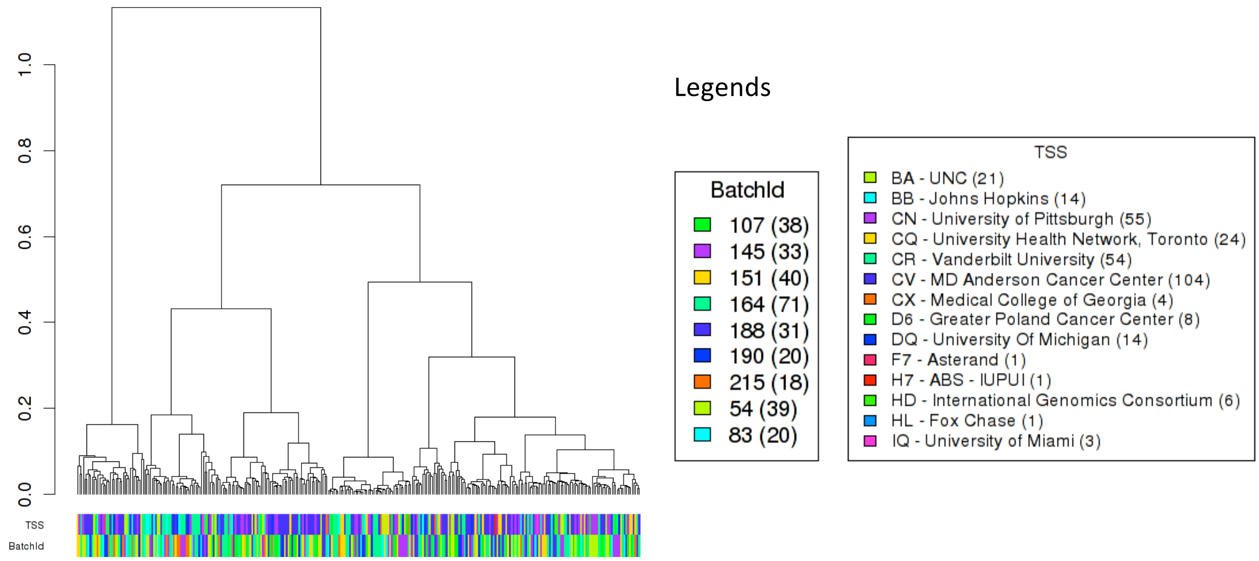


Figure S10.4 Hierarchical clustering plot for DNA methylation HM450 data.

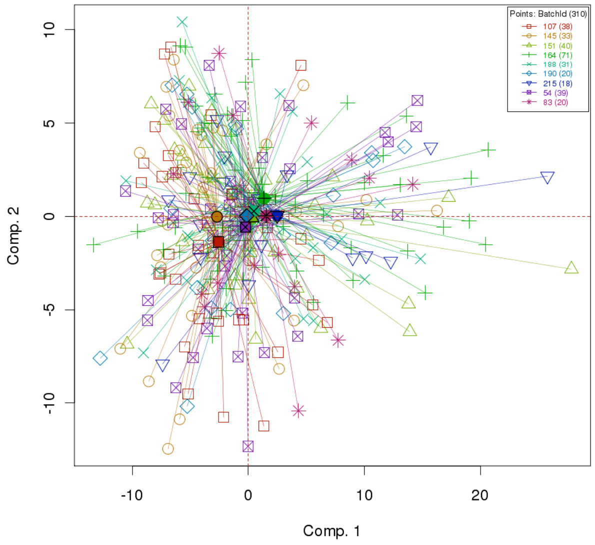


Figure S10.5 PCA for DNA methylation, with samples connected by centroids according to batch ID.

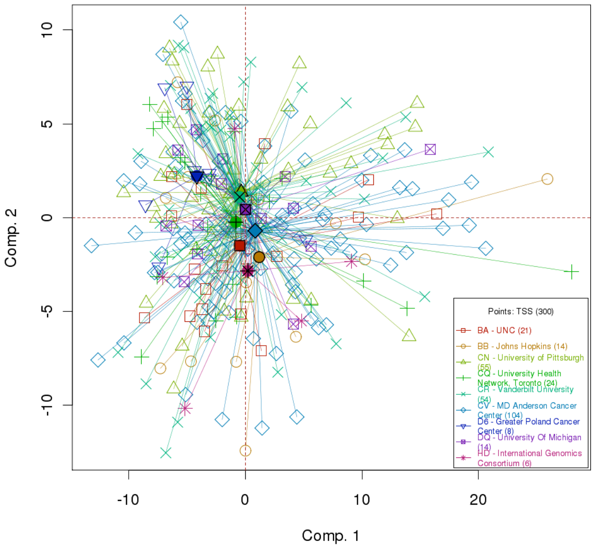


Figure S10.6 PCA for DNA methylation, with samples connected by centroids according to tissue source site.

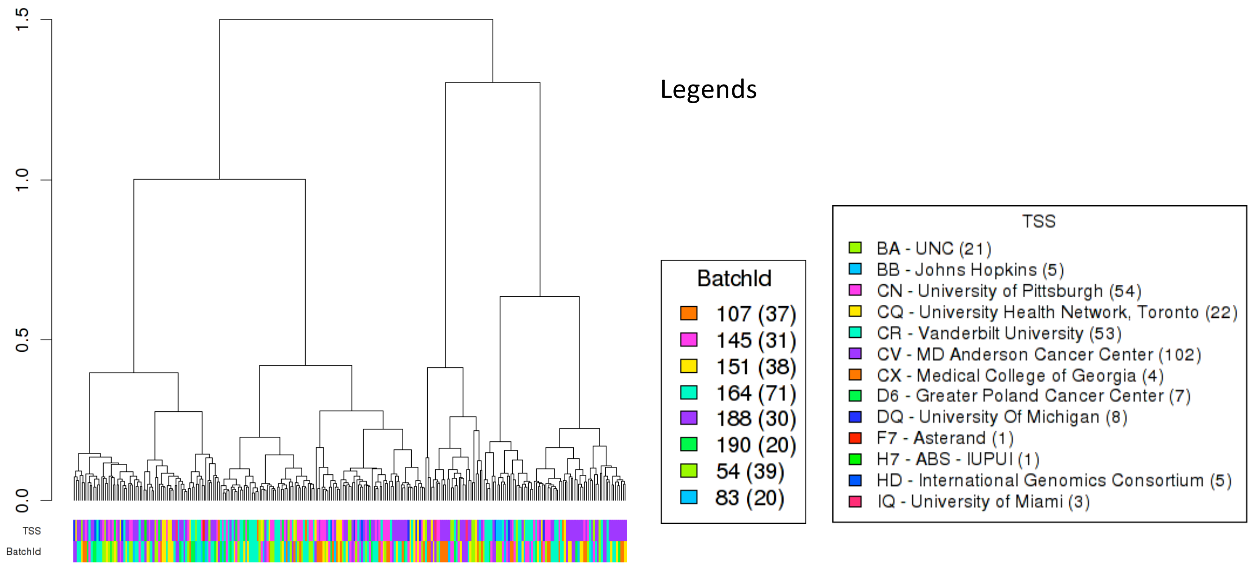


Figure S10.7 Hierarchical clustering for mRNA expression from RNA-Seq data.

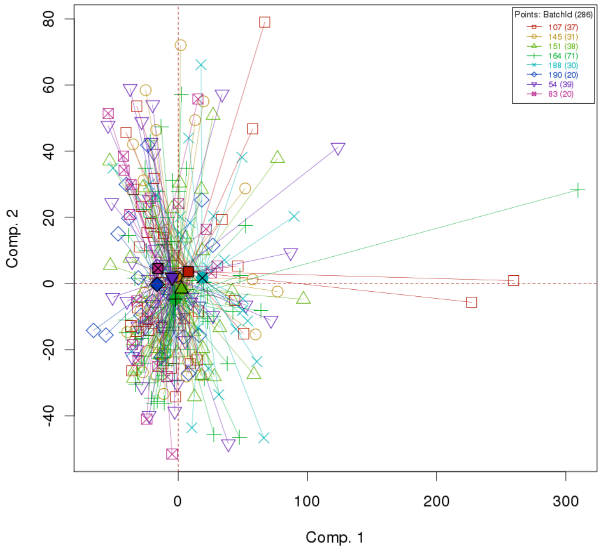


Figure S10.8 PCA: First two principal components for RNA-Seq, with samples connected by centroids according to batch ID.

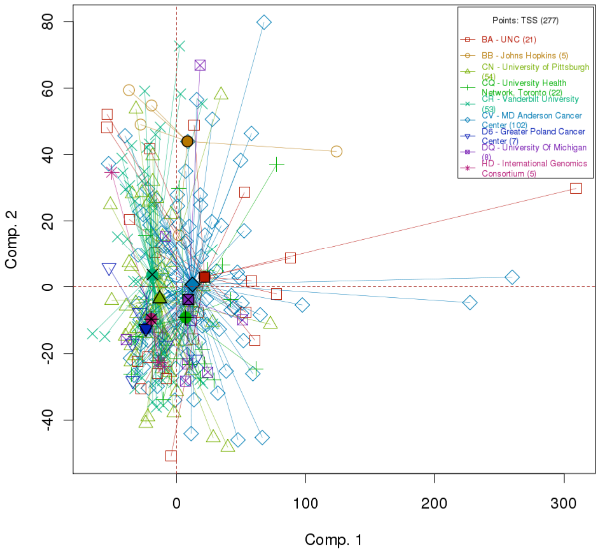


Figure S10.9 PCA: First two principal components for RNA-Seq, with samples connected by centroids according to tissue source site.

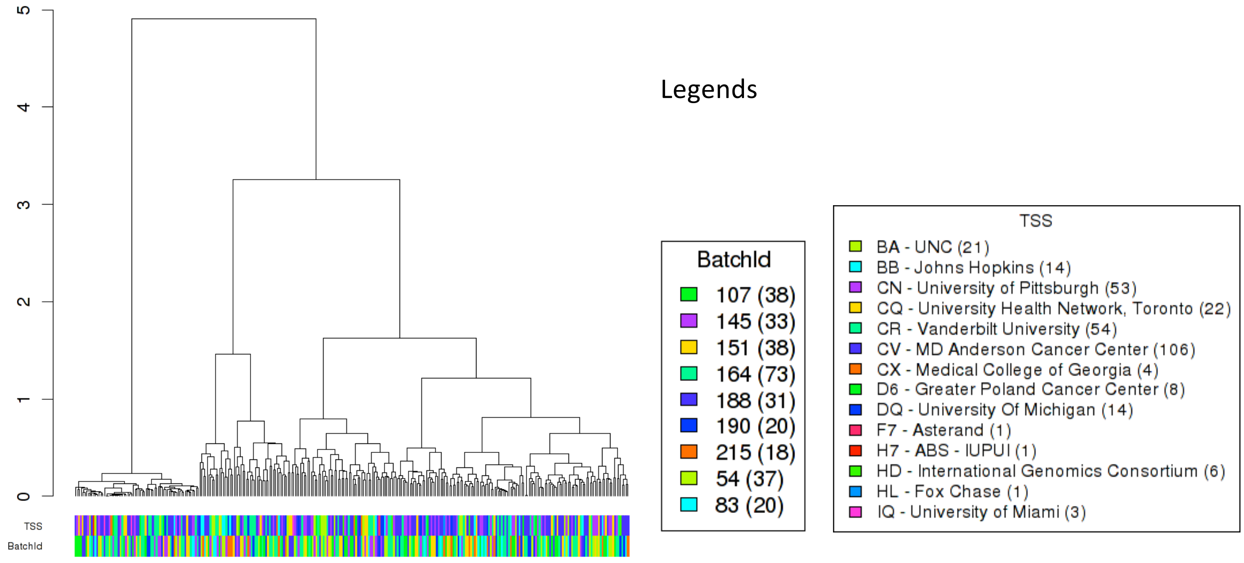
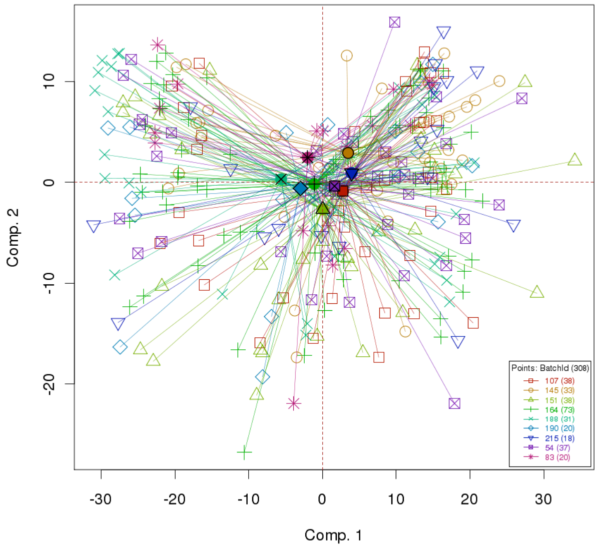


Figure S10.10 Hierarchical clustering for SNP6 data.



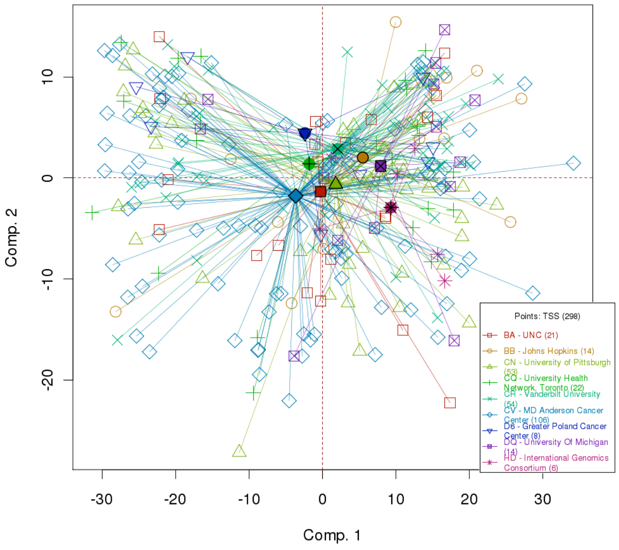


Figure S10.11 PCA: First two principal components for SNP6, with samples connected by centroids according to batch ID.

Figure S10.12 PCA: First two principal components for SNP6, with samples connected by centroids according to tissue source site.

1 Ang, K. K. *et al.* Head and neck carcinoma in the United States: first comprehensive report of the Longitudinal Oncology Registry of Head and Neck Carcinoma (LORHAN). *Cancer* **118**, 5783-5792, doi:10.1002/cncr.27609 (2012).

2 Edge, S. B., American Joint Committee on Cancer. & American Cancer Society. *AJCC cancer staging handbook: from the AJCC cancer staging manual*. 7th edn, (2010).

3 Stadler, M. E., Patel, M. R., Couch, M. E. & Hayes, D. N. Molecular biology of head and neck cancer: risks and pathways. *Hematology/oncology clinics of North America* **22**, 1099-1124, vii, doi:10.1016/j.hoc.2008.08.007 (2008).

4 Janot, F. *et al.* Randomized trial of postoperative reirradiation combined with chemotherapy after salvage surgery compared with salvage surgery alone in head and neck carcinoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **26**, 5518-5523, doi:10.1200/JCO.2007.15.0102 (2008).

5 Ang, K. K. *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *The New England journal of medicine* **363**, 24-35, doi:10.1056/NEJMoa0912217 (2010).

6 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178, doi:10.1093/nar/gkq622 (2010).

7 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

8 R: a language and environment for statistical computing. http://www.R-project.org. R Core Team. R Foundation for Statistical Computing (Vienna, Austria, 2014).

9 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160, doi:10.1126/science.1208130 (2011).

10 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

11 Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic acids research* **41**, D571-578, doi:10.1093/nar/gks984 (2013).

12 *The NCBI handbook [Internet]*, <http://www.ncbi.nlm.nih.gov/books/NBK21101> (2002).

13 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

14 Gao, G. *et al.* A microRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer* **119**, 72-80, doi:10.1002/cncr.27696 (2013).

15 Hui, A. B. *et al.* Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 2154-2162, doi:10.1158/1078-0432.CCR-12-3572 (2013).

16 Lajer, C. B. *et al.* The role of miRNAs in human papilloma virus (HPV)-associated cancers: bridging between HPV-related head and neck cancer and cervical cancer. *British journal of cancer* **106**, 1526-1534, doi:10.1038/bjc.2012.109 (2012).

17 Lajer, C. B. *et al.* Different miRNA signatures of oral and pharyngeal squamous cell carcinomas: a prospective translational study. *British journal of cancer* **104**, 830-840, doi:10.1038/bjc.2011.29 (2011).

18 Limma: linear models for microarray data (Springer-Verlag, New York, 2005).

19 Lechner, M. *et al.* Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. *Genome medicine* **5**, 15, doi:10.1186/gm419 (2013).

20 Walline, H. M. *et al.* High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and oral cavity cancers: comparison of multiple methods. *JAMA otolaryngology-- head & neck surgery* **139**, 1320-1327, doi:10.1001/jamaoto.2013.5460 (2013).

21 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

22 Therneau, T. M. *A package for survival analysis in S, R package version 2.37-7*, <http://CRAN.R-project.org/package=survival> (2014).

23 Smeets, S. J. *et al.* Genetic classification of oral and oropharyngeal carcinomas identifies subgroups with a different prognosis. *Cellular oncology : the official journal of the International Society for Cellular Oncology* **31**, 291-300, doi:10.3233/CLO-2009-0471 (2009).

24 McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* **40**, 1166-1174, doi:10.1038/ng.238 (2008).

25 Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* **40**, 1253-1260, doi:10.1038/ng.237 (2008).

26 Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

27 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572, doi:10.1093/biostatistics/kxh008 (2004).

28 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

29 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

30 <http://gdac.broadinstitute.org/runs/analyses__2013_02_22/reports/index.html> (2013).

31 Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

32 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

33 *http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf*.

34 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

35 *https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/luad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUAD.IlluminaHiSeq_RNASeqV2.mage-tab.1.9.0/DESCRIPTION.txt*.

36 https://cghub.ucsc.edu/docs/tcga/UNC_mRNAseq_summary.pdf.

37 *https://tcga-data.nci.nih.gov/docs/publications/hnsc_2014*.

38 Wilkerson MD, C. C., Sun W, Hoadley KA, Walter V, Mose LE, Troester MA, Hammerman PS, Parker JS, Perou CM, Hayes DN. Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic acids research* (2014).

39 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

40 Kimes, P. K. *et al.* SigFuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. *Nucleic acids research* (2014).

41 Talieri, M. *et al.* Human kallikrein-related peptidase 12 (KLK12) splice variants expression in breast cancer and their clinical impact. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **33**, 1075-1084, doi:10.1007/s13277-012-0347-x (2012).

42 Wu, Y. M. *et al.* Identification of targetable FGFR gene fusions in diverse cancers. *Cancer discovery* **3**, 636-647, doi:10.1158/2159-8290.CD-13-0050 (2013).

43 Reich, M. *et al.* GenePattern 2.0. *Nature genetics* **38**, 500-501, doi:10.1038/ng0506-500 (2006).

44 Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome research* **23**, 228-235, doi:10.1101/gr.141382.112 (2013).

45 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).

46 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

47 Futreal, P. A. *et al.* A census of human cancer genes. *Nature reviews. Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

48 Cohen, J. *et al.* Attenuated transforming growth factor beta signaling promotes nuclear factor-kappaB activation in head and neck cancer. *Cancer research* **69**, 3415-3424, doi:10.1158/0008-5472.CAN-08-3704 (2009).

49 Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

50 Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1128-1136, doi:10.1073/pnas.1110574108 (2011).

51 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

52 Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929, doi:10.1016/j.cell.2013.04.010 (2013).

53 K. Pruitt, G. B., T. Tatusova, D. Maglott. The Reference Sequence (RefSeq) Database. *The NCBI Handbook [Internet]* (2002).

54 Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40**, D290-301, doi:10.1093/nar/gkr1065 (2012).

55 Quintana, R. M. *et al.* A transposon-based analysis of gene mutations related to skin cancer development. *The Journal of investigative dermatology* **133**, 239-248, doi:10.1038/jid.2012.245 (2013).

56 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).

57 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1, doi:10.1126/scisignal.2004088 (2013).

58 Van Waes, C. *et al.* Molecular and clinical responses in a pilot study of gefitinib with paclitaxel and radiation in locally advanced head-and-neck cancer. *International journal of radiation oncology, biology, physics* **77**, 447-454, doi:10.1016/j.ijrobp.2009.05.037 (2010).

59 Chung, C. H. *et al.* Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer cell* **5**, 489-500 (2004).

60 Walter, V. *et al.* Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PloS one* **8**, e56823, doi:10.1371/journal.pone.0056823 (2013).

61 Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell* **17**, 98-110, doi:10.1016/j.ccr.2009.12.020 (2010).

62 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121, doi:10.1073/pnas.091062498 (2001).

63 Slebos, R. J. *et al.* Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* **12**, 701-709, doi:10.1158/1078-0432.CCR-05-2017 (2006).

64 Schlecht, N. F. *et al.* Gene expression profiles in HPV-infected head and neck cancer. *The Journal of pathology* **213**, 283-293, doi:10.1002/path.2227 (2007).

65 Lohavanichbutr, P. *et al.* Genomewide gene expression profiles of HPV-positive and HPV-negative oropharyngeal cancer: potential implications for treatment choices. *Archives of otolaryngology--head & neck surgery* **135**, 180-188, doi:10.1001/archoto.2008.540 (2009).

66 Heibein, A. D., Guo, B., Sprowl, J. A., Maclean, D. A. & Parissenti, A. M. Role of aldo-keto reductases and other doxorubicin pharmacokinetic genes in doxorubicin resistance, DNA binding, and subcellular localization. *BMC cancer* **12**, 381, doi:10.1186/1471-2407-12-381 (2012).

67 Marcato, P., Dean, C. A., Giacomantonio, C. A. & Lee, P. W. Aldehyde dehydrogenase: its role as a cancer stem cell marker comes down to the specific isoform. *Cell Cycle* **10**, 1378-1384 (2011).

68 Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. *Nature reviews. Cancer* **10**, 116-129, doi:10.1038/nrc2780 (2010).

69 Huang, L., Walter, V., Hayes, D. N. & Onaitis, M. Hedgehog-GLI signaling inhibition suppresses tumor growth in squamous lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **20**, 1566-1575, doi:10.1158/1078-0432.CCR-13-2195 (2014).

70 Bass, A. J. *et al.* SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature genetics* **41**, 1238-1242, doi:10.1038/ng.465 (2009).

71 Salama, I., Malone, P. S., Mihaimeed, F. & Jones, J. L. A review of the S100 proteins in cancer. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* **34**, 357-364, doi:10.1016/j.ejso.2007.04.009 (2008).

72 Karantza, V. Keratins in health and cancer: more than mere epithelial cell markers. *Oncogene* **30**, 127-138, doi:10.1038/onc.2010.456 (2011).

73 Chen, J. *et al.* Significance of CD44 expression in head and neck cancer: a systemic review and meta-analysis. *BMC cancer* **14**, 15, doi:10.1186/1471-2407-14-15 (2014).

74 Sparano, A. *et al.* Genome-wide profiling of oral squamous cell carcinoma by array-based comparative genomic hybridization. *The Laryngoscope* **116**, 735-741, doi:10.1097/01.mlg.0000205141.54471.7f (2006).

75 Yu, Y. H., Kuo, H. K. & Chang, K. W. The evolving transcriptome of head and neck squamous cell carcinoma: a systematic review. *PloS one* **3**, e3215, doi:10.1371/journal.pone.0003215 (2008).

76 Gombos, K. *et al.* miRNA expression profiles of oral squamous cell carcinomas. *Anticancer research* **33**, 1511-1517 (2013).

77 Hui, A. B. *et al.* Significance of dysregulated metadherin and microRNA-375 in head and neck cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 7539-7550, doi:10.1158/1078-0432.CCR-11-2102 (2011).

78 Hui, A. B. *et al.* Comprehensive MicroRNA profiling for head and neck squamous cell carcinomas. *Clinical cancer research : an official journal of the American Association for Cancer Research* **16**, 1129-1139, doi:10.1158/1078-0432.CCR-09-2166 (2010).

79 Nohata, N. *et al.* Tumor suppressive microRNA-375 regulates oncogene AEG-1/MTDH in head and neck squamous cell carcinoma (HNSCC). *Journal of human genetics* **56**, 595-601, doi:10.1038/jhg.2011.66 (2011).

80 Christensen, B. C. *et al.* Mature microRNA sequence polymorphism in MIR196A2 is associated with risk and prognosis of head and neck cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **16**, 3713-3720, doi:10.1158/1078-0432.CCR-10-0657 (2010).

81 Cervigne, N. K. *et al.* Identification of a microRNA signature associated with progression of leukoplakia to oral carcinoma. *Human molecular genetics* **18**, 4818-4829, doi:10.1093/hmg/ddp446 (2009).

82 Lenarduzzi, M. *et al.* MicroRNA-193b enhances tumor progression via down regulation of neurofibromin 1. *PloS one* **8**, e53765, doi:10.1371/journal.pone.0053765 (2013).

83 Chen, Z. *et al.* Down-regulation of the microRNA-99 family members in head and neck squamous cell carcinoma. *Oral oncology* **48**, 686-691, doi:10.1016/j.oraloncology.2012.02.020 (2012).

84 Wilkerson, M. D. *et al.* Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical cancer research : an official journal of the American Association for Cancer Research* **16**, 4864-4875, doi:10.1158/1078-0432.CCR-10-0199 (2010).

85 Tibes, R. *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics* **5**, 2512-2521, doi:10.1158/1535-7163.MCT-06-0334 (2006).

86 Liang, J. *et al.* The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. *Nature cell biology* **9**, 218-224, doi:10.1038/ncb1537 (2007).

87 Hu, J. *et al.* Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986-1994, doi:10.1093/bioinformatics/btm283 (2007).

88 Hennessy, B. T. *et al.* Pharmacodynamic markers of perifosine efficacy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 7421-7431, doi:10.1158/1078-0432.CCR-07-0760 (2007).

89 SuperCurve: SuperCurve Package. R package v. 1.4.1 (2011).

90 Gonzalez-Angulo, A. M. *et al.* Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clinical proteomics* **8**, 11, doi:10.1186/1559-0275-8-11 (2011).

91 Hennessy, B. T. *et al.* A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clinical proteomics* **6**, 129-151, doi:10.1007/s12014-010-9055-y (2010).

92 Ciriello, G., Cerami, E., Aksoy, B. A., Sander, C. & Schultz, N. Using MEMo to discover mutual exclusivity modules in cancer. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 8**, Unit 8 17, doi:10.1002/0471250953.bi0817s41 (2013).

93 Ehsanian, R. *et al.* YAP dysregulation by phosphorylation or DeltaNp63-mediated gene repression promotes proliferation, survival and migration in head and neck cancer subsets. *Oncogene* **29**, 6160-6171, doi:10.1038/onc.2010.339 (2010).

94 Brahmer, J. R. *et al.* Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 3167-3175, doi:10.1200/JCO.2009.26.7609 (2010).

95 Lynch, T. J. *et al.* Ipilimumab in combination with paclitaxel and carboplatin as first-line treatment in stage IIIB/IV non-small-cell lung cancer: results from a randomized, double-blind, multicenter phase II study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **30**, 2046-2054, doi:10.1200/JCO.2011.38.4032 (2012).

96 Smeets, S. J. *et al.* Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus. *Oncogene* **25**, 2558-2564, doi:10.1038/sj.onc.1209275 (2006).

97 Poage, G. M. *et al.* Identification of an epigenetic profile classifier that is associated with survival in head and neck cancer. *Cancer research* **72**, 2728-2737, doi:10.1158/0008-5472.CAN-11-4121-T (2012).

98 Byers, L. A. *et al.* An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 279-290, doi:10.1158/1078-0432.CCR-12-1558 (2013).

99 Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

100 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**, 367, doi:10.1186/1471-2105-11-367 (2010).

101 *http://bonsai.hgc.jp/~mdehoon/software/cluster/*.

102 *jtreeview.sourceforge.net/*.

103 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573, doi:10.1093/bioinformatics/btq170 (2010).

104 Gregory, P. A. *et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology* **10**, 593-601, doi:10.1038/ncb1722 (2008).

105 Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007-20012, doi:10.1073/pnas.0710052104 (2007).

106 Li, X., Gill, R., Cooper, N. G., Yoo, J. K. & Datta, S. Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. *BMC medical genomics* **4**, 44, doi:10.1186/1755-8794-4-44 (2011).

107 Buffa, F. M. *et al.* microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer research* **71**, 5635-5645, doi:10.1158/0008-5472.CAN-11-0489 (2011).

108 Huang, J. C. *et al.* Using expression profiling data to identify human microRNA targets. *Nature methods* **4**, 1045-1049, doi:10.1038/nmeth1130 (2007).

109 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:10.1093/bioinformatics/bts163 (2012).

110 Hsu, S. D. *et al.* miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic acids research* **39**, D163-169, doi:10.1093/nar/gkq1107 (2011).

111 Wong, J. V., Dong, P., Nevins, J. R., Mathey-Prevot, B. & You, L. Network calisthenics: control of E2F dynamics in cell cycle entry. *Cell Cycle* **10**, 3086-3094 (2011).

112 Zhu, M. *et al.* MISP is a novel Plk1 substrate required for proper spindle orientation and mitotic progression. *The Journal of cell biology* **200**, 773-787, doi:10.1083/jcb.201207050 (2013).

113 Tan, E. J. *et al.* Regulation of transcription factor Twist expression by the DNA architectural protein high mobility group A2 during epithelial-to-mesenchymal transition. *The Journal of biological chemistry* **287**, 7134-7145, doi:10.1074/jbc.M111.291385 (2012).

114 Sun, M. *et al.* HMGA2/TET1/HOXA9 signaling pathway regulates breast cancer growth and metastasis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 9920-9925, doi:10.1073/pnas.1305172110 (2013).

115 Morris, L. G. *et al.* Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nature genetics* **45**, 253-261, doi:10.1038/ng.2538 (2013).

116 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).

117 Hayward, P., Kalmar, T. & Arias, A. M. Wnt/Notch signalling and information processing during development. *Development* **135**, 411-424, doi:10.1242/dev.000505 (2008).

118 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245, doi:10.1093/bioinformatics/btq182 (2010).

119 Sedgewick, A. J., Benz, S. C., Rabizadeh, S., Soon-Shiong, P. & Vaske, C. J. Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. *Bioinformatics* **29**, i62-70, doi:10.1093/bioinformatics/btt229 (2013).

120 Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640-i646, doi:10.1093/bioinformatics/bts402 (2012).

121 Wong, C. K. *et al.* The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic acids research* **41**, W218-224, doi:10.1093/nar/gkt473 (2013).

122 Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154-1157, doi:10.1126/science.1206923 (2011).

123 Shibata, T. *et al.* Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13568-13573, doi:10.1073/pnas.0806268105 (2008).

124 Bancroft, C. C. *et al.* Effects of pharmacologic antagonists of epidermal growth factor receptor, PI3K and MEK signal kinases on NF-kappaB and AP-1 activation and IL-8 and VEGF expression in human

head and neck squamous cell carcinoma lines. *International journal of cancer. Journal international du cancer* **99**, 538-548, doi:10.1002/ijc.10398 (2002).

125 Herzog, A. *et al.* PI3K/mTOR inhibitor PF-04691502 antitumor activity is enhanced with induction of wild-type TP53 in human xenograft and murine knockout models of head and neck cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 3808-3819, doi:10.1158/1078-0432.CCR-12-2716 (2013).

126 Lui, V. W. *et al.* Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer discovery* **3**, 761-769, doi:10.1158/2159-8290.CD-13-0103 (2013).

127 Leonard, J. P. *et al.* Selective CDK4/6 inhibition with tumor responses by PD0332991 in patients with mantle cell lymphoma. *Blood* **119**, 4597-4607, doi:10.1182/blood-2011-10-388298 (2012).

128 Chung, C. H. & Gillison, M. L. Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. *Clinical cancer research : an official journal of the American Association for Cancer Research* **15**, 6758-6762, doi:10.1158/1078-0432.CCR-09-0784 (2009).

129 Singh, D. *et al.* Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science* **337**, 1231-1235, doi:10.1126/science.1220834 (2012).

130 Sok, J. C. *et al.* Mutant epidermal growth factor receptor (EGFRvIII) contributes to head and neck cancer growth and resistance to EGFR targeting. *Clinical cancer research : an official journal of the American Association for Cancer Research* **12**, 5064-5073, doi:10.1158/1078-0432.CCR-06-0913 (2006).

131 Knoll, S., Emmrich, S. & Putzer, B. M. The E2F1-miRNA cancer progression network. *Advances in experimental medicine and biology* **774**, 135-147, doi:10.1007/978-94-007-5590-1_8 (2013).

132 Mayo, M. W. *et al.* Requirement of NF-kappaB activation to suppress p53-independent apoptosis induced by oncogenic Ras. *Science* **278**, 1812-1815 (1997).

133 Oberst, A. & Green, D. R. It cuts both ways: reconciling the dual roles of caspase 8 in cell death and survival. *Nature reviews. Molecular cell biology* **12**, 757-763, doi:10.1038/nrm3214 (2011).

134 Chen, G. *et al.* Phosphorylated FADD induces NF-kappaB, perturbs cell cycle, and is associated with poor outcome in lung adenocarcinomas. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12507-12512, doi:10.1073/pnas.0500397102 (2005).

135 Van Waes, C. Nuclear factor-kappaB in development, prevention, and therapy of cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 1076-1082, doi:10.1158/1078-0432.CCR-06-2221 (2007).

136 Liebowitz, D. Epstein-Barr virus and a cellular signaling pathway in lymphomas from immunosuppressed patients. *The New England journal of medicine* **338**, 1413-1421, doi:10.1056/NEJM199805143382003 (1998).

137 Oganesyan, G. *et al.* Critical role of TRAF3 in the Toll-like receptor-dependent and -independent antiviral response. *Nature* **439**, 208-211, doi:10.1038/nature04374 (2006).

138 Eliopoulos, A. G. *et al.* CD40-induced growth inhibition in epithelial cells is mimicked by Epstein-Barr Virus-encoded LMP1: involvement of TRAF3 as a common mediator. *Oncogene* **13**, 2243-2254 (1996).

139 Imbeault, M. *et al.* Acquisition of host-derived CD40L by HIV-1 in vivo and its functional consequences in the B-cell compartment. *Journal of virology* **85**, 2189-2200, doi:10.1128/JVI.01993-10 (2011).

140 Chung, G. T. *et al.* Constitutive activation of distinct NF-kappaB signals in EBV-associated nasopharyngeal carcinoma. *The Journal of pathology* **231**, 311-322, doi:10.1002/path.4239 (2013).

141 Annunziata, C. M. *et al.* Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. *Cancer cell* **12**, 115-130, doi:10.1016/j.ccr.2007.07.004 (2007).

142 Hacker, H., Tseng, P. H. & Karin, M. Expanding TRAF function: TRAF3 as a tri-faced immune regulator. *Nature reviews. Immunology* **11**, 457-468, doi:10.1038/nri2998 (2011).

143 Esteban, F. *et al.* Lack of MHC class I antigens and tumour aggressiveness of the squamous cell carcinoma of the larynx. *British journal of cancer* **62**, 1047-1051 (1990).

144 Keating, P. J. *et al.* Frequency of down-regulation of individual HLA-A and -B alleles in cervical carcinomas in relation to TAP-1 expression. *British journal of cancer* **72**, 405-411 (1995).

145 Yang, X. *et al.* DeltaNp63 versatilely regulates a Broad NF-kappaB gene program and promotes squamous epithelial proliferation, migration, and inflammation. *Cancer research* **71**, 3688-3700, doi:10.1158/0008-5472.CAN-10-3445 (2011).

146 Lu, H. *et al.* TNF-alpha promotes c-REL/DeltaNp63alpha interaction and TAp73 dissociation from key genes that mediate growth arrest and apoptosis in head and neck cancer. *Cancer research* **71**, 6867-6877, doi:10.1158/0008-5472.CAN-11-2460 (2011).

147 Nguyen, B. C. *et al.* Cross-regulation between Notch and p63 in keratinocyte commitment to differentiation. *Genes & development* **20**, 1028-1042, doi:10.1101/gad.1406006 (2006).

148 Dotto, G. P. Crosstalk of Notch with p53 and p63 in cancer growth control. *Nature reviews. Cancer* **9**, 587-595, doi:10.1038/nrc2675 (2009).

149 Hast, B. E. *et al.* Cancer-derived mutations in KEAP1 impair NRF2 degradation but not ubiquitination. *Cancer research* **74**, 808-817, doi:10.1158/0008-5472.CAN-13-1655 (2014).

150 Haraguchi, K. *et al.* Ajuba negatively regulates the Wnt signaling pathway by promoting GSK-3beta-mediated phosphorylation of beta-catenin. *Oncogene* **27**, 274-284, doi:10.1038/sj.onc.1210644 (2008).

151 Nagai, Y. *et al.* The LIM protein Ajuba is required for ciliogenesis and left-right axis determination in medaka. *Biochemical and biophysical research communications* **396**, 887-893, doi:10.1016/j.bbrc.2010.05.017 (2010).

152 Sun, G. & Irvine, K. D. Ajuba family proteins link JNK to Hippo signaling. *Science signaling* **6**, ra81, doi:10.1126/scisignal.2004324 (2013).

153 Kalan, S., Matveyenko, A. & Loayza, D. LIM Protein Ajuba Participates in the Repression of the ATR-Mediated DNA Damage Response. *Frontiers in genetics* **4**, 95, doi:10.3389/fgene.2013.00095 (2013).

154 Moriyama, M. *et al.* Multiple roles of Notch signaling in the regulation of epidermal development. *Developmental cell* **14**, 594-604, doi:10.1016/j.devcel.2008.01.017 (2008).

155 Wei, Y. *et al.* Nrf2 acts cell-autonomously in endothelium to regulate tip cell formation and vascular branching. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E3910-3918, doi:10.1073/pnas.1309276110 (2013).

156 Becks, L. *et al.* Aggressive mammary carcinoma progression in Nrf2 knockout mice treated with 7,12-dimethylbenz[a]anthracene. *BMC cancer* **10**, 540, doi:10.1186/1471-2407-10-540 (2010).

157 Cohen, E. E. Role of epidermal growth factor receptor pathway-targeted therapy in patients with recurrent and/or metastatic squamous cell carcinoma of the head and neck. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **24**, 2659-2665, doi:10.1200/JCO.2005.05.4577 (2006).

158 D'Amato, N. C., Howe, E. N. & Richer, J. K. MicroRNA regulation of epithelial plasticity in cancer. *Cancer letters* **341**, 46-55, doi:10.1016/j.canlet.2012.11.054 (2013).

159 Johnson, C. D. *et al.* The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer research* **67**, 7713-7722, doi:10.1158/0008-5472.CAN-07-1083 (2007).

160 Hebert, S. S. *et al.* MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. *Neurobiology of disease* **33**, 422-428, doi:10.1016/j.nbd.2008.11.009 (2009).

161 Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295, doi:10.1016/j.ygeno.2011.07.007 (2011).

162 Bibikova, M. & Fan, J. B. Genome-wide DNA methylation profiling. *Wiley interdisciplinary reviews. Systems biology and medicine* **2**, 210-223, doi:10.1002/wsbm.35 (2010).

163 Davis, S., Du, P., Bilke, S., Triche Jr., T., Bootwalla, M. *methylumi: Handle Illumina methylation data. R package version 2.0.4.*, (2012).