doi:10.1038/nature21386

S1. Biospecimen Collection and Clinical Data

Sample Acquisition

Resection and biopsy biospecimens were collected from patients diagnosed with cervical squamous cell carcinoma, endocervical adenocarcinoma, or adenosquamous carcinoma that had not received prior chemotherapy or radiotherapy. Institutional review boards at each tissue source site (TSS) reviewed protocols and consent documentation and approved submission of cases to TCGA. Cases were staged according to the American Joint Committee on Cancer (AJCC) and International Federation of Gynecology and Obstetrics (FIGO) staging systems. Each frozen primary tumor specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the TSS). Normal uterus was also submitted for some cases. Specimens were shipped overnight from 20 TSSs using a cryoport that maintained an average temperature of less than -180°C.

Pathology quality control was performed on each tumor and adjacent normal tissue (if available) specimen from either a frozen section slide prepared by the Biospecimen Core Resource (BCR) or from a frozen section slide prepared by the TSS. Hematoxylin and eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumor specimen was histologically consistent with the allowable cervical cancers and the adjacent normal specimen contained no tumor cells. The percent tumor nuclei, percent necrosis, and other pathology annotations were also assessed. Tumor samples with ≥60% tumor nuclei and ≤20% necrosis were submitted for nucleic acid extraction.

Approximately 61% of cervical cancer cases (consisting of a primary tumor and a germline control) submitted to the BCR and processed passed quality control metrics. Tumor tissue from 173 cases was submitted for reverse phase protein array (RPPA) analysis.

TSSs contributing biospecimens included: Analytical Biological Services, Inc., Asterand, Inc., Barretos Cancer Hospital, Baylor College of Medicine, Candler, Catholic Health Initiative - Penrose St. Francis Health Services, Cedars-Sinai Medical Center, Christiana Care Health Services, Inc., Gynecologic Oncology Group, Indiana University School of Medicine, International Genomics Consortium, ILSbio, LLC., The University of Texas MD Anderson Cancer Center, Medical College of Wisconsin, Montefiore Medical Center, Memorial Sloan Kettering Cancer Center, National Cancer Institute, Ontario Tumour Bank – London Health Sciences Centre, Ontario Institute for Cancer Research – Ottawa, ProteoGenex, Roswell Park Cancer Institute, University of Hawaii, University of Kansas, University of Minnesota, University of New Mexico, University of North Carolina, University of Oklahoma Health Sciences Center, University of Pittsburgh, University of Washington, and Washington University in St. Louis.

Sample Processing

DNA and RNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

RNA samples were quantified by measuring Abs$_{260}$ with a UV spectrophotometer and DNA was quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify that tumor DNA and germline DNA representing a case were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen (Hilden, Germany) for REPLI-g whole genome amplification using a 100 μg reaction scale. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only analytes with RIN ≥7.0 were included in this study. Only cases yielding a minimum of 6.9 μg of tumor DNA, 5.15 μg RNA, and 4.9 μg of germline DNA were included in this study.

Samples with residual tumor tissue were considered for proteomics analysis. When available, a 10-20 mg piece of snap-frozen tumor adjacent to the piece used for molecular sequencing and characterization was submitted to MD Anderson Cancer Center for RPPA analysis.

Data Freeze

Details of the data freeze samples are described in Methods. Overall, data from 228 samples was used in various analyses across six different clinical and molecular platforms, which comprises the largest cervical cancer dataset to date (Extended Data Fig. 1a).

Histology Verification

Frozen sections of all cervical cancers submitted for TCGA analysis were reviewed by a tissue site pathologist and an independent pathologist prior to acceptance into the study. When available, scanned images of the formalin-fixed, paraffin embedded tissue slides were reviewed by an expert pathology panel. Only cases that met criteria for primary cervical cancer according to WHO criteria[68] were accepted. These included squamous cell carcinomas, both large cell keratinizing in which at least one well-formed keratin pearl was identified and large cell non-keratinizing. Adenocarcinomas included adenocarcinoma of usual type, including mucin depleted, mucinous, and endometrioid type. For analysis purposes, all adenocarcinomas were combined into one endocervical adenocarcinoma category. Three adenosquamous carcinomas were also included. All cervical cancers were assigned a pathologic grade, including Grade I: well-differentiated; Grade II: moderately differentiated; and Grade III: poorly differentiated. Care was taken to verify that the tumors included were not endometrial in origin.

S2. HPV Detection and Integration

HPV Detection by MassArray (Nationwide Children's Hospital)

HPV status was determined by an ultra-sensitive method using real-time competitive polymerase chain reaction and matrix-assisted laser desorption/ionization-time of flight mass spectroscopy with separation of products on a matrix-loaded silicon chip array, similar to the work described in Tang *et al*[45]. Multiplex PCR amplification of the E6 region of 16 discrete high-risk HPV types (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 90), 2 low-risk HPV types (HPV 6 and 11), and human GAPDH control was run to saturation followed by shrimp alkaline phosphatase quenching. Amplification reactions included a competitor oligo identical to each natural amplicon except for a single nucleotide difference. Probes that identify unique sequences in the oncogenic E6 region of each type were used in multiplex single base extension reactions extending at the single base difference between wild-type and competitor HPV so that each HPV type and its competitor were distinguished by mass when analyzed on the MALDI-TOF mass spectrometer.

Pathogen Detection from RNA-seq Data by BioBloom Tools (BC Cancer Agency)

The microbial detection pipeline used by the BC Cancer Agency's Genome Sciences Centre (BC) is based on BioBloom Tools (BBT, v1.2.4b1), which is a Bloom filter-based method for rapidly classifying RNA-seq or DNA-seq read sequences[46]. We generated 43 filters from "complete" NCBI genome reference sequences of bacteria, viruses, fungi, and protozoa, using 25-bp k-mers and a false positive rate of 0.02. We ran BBT in paired-end (PE) mode with a sliding window to screen FASTQ files from RNA-seq libraries (48-bp PE reads, 178 tumors and no adjacent tissue normals), and 40 whole genome shotgun libraries (WGS, 50-bp PE reads, 19 tumors and 19 blood normals). In a single-pass scan for each library, BBT categorized each read pair as matching the human filter, matching a unique microbial filter, matching more than one filter (multi-match), or matching neither human nor microbe (no-match). For each filter, we then calculated a reads-per-million (RPM) abundance metric as:

$$Abundance\ metric = \left(\frac{\#reads\ mapped\ to\ a\ microbe\ filter}{\#chastity\ passed\ reads\ in\ the\ sample} * 10^6 \right)$$

HPV-specific detection thresholds were identified from distinct gaps between HPV-positive and HPV-negative libraries in sorted RPM profiles. For HPV, we applied thresholds of 1.8 and 0.4 RPM to RNA-seq and WGS profiles, respectively. Of note, different microbes may require different thresholds. To identify the specific HPV strain in each positive library, we scanned the reads that had been classified as HPV against separate filters for each of the reference HPV strains, using single-pass BBT runs. The classified FASTQ files were then passed into the viral integration analysis stage (below).

Pathogen Detection from RNA-seq Data by PathSeq (Broad Institute)

The PathSeq algorithm[47] 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 to HPV genomes in RNA sequencing data.

Subjects were classified as HPV-positive by RNA sequencing if at least 1 HPV read in 1 million human reads were present; otherwise, subjects were classified as HPV-negative. Using PathSeq, human reads were subtracted by first mapping reads to a database of human genomes using BWA (version 0.6.1)[69], Megablast (version 2.2.23), and Blastn (version 2.2.23)[70]. 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 with Megablast to a database of microbial genomes that includes multiple HPV reference genomes. HPV reference genomes were obtained from the NCBI nucleotide database (downloaded in June 2013).

Pathogen Detection from Low-Pass WGS Data (Harvard Medical School)

An in-house developed pipeline, PathWatch, was used to detect bacteria and viruses and to examine the integration status of the bacterial/viral genome. First, computational subtraction of sequences mapped previously to the human genome was performed. Next, BWA was used to map the remaining set of non-human sequences to the set of bacterial and viral reference genomes obtained from the NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/release/microbial/ and ftp://ftp.ncbi.nih.gov/refseq/release/viral/ respectively). Reads that aligned to the genomes of multiple species were filtered out. The percentage of covered pathogen genome, count of pathogen sequencing reads normalized by the length of the pathogen genome, and total number of non-human reads in the sample were calculated. To consider a given sample positive for the pathogen presence we chose an empirical threshold of 1 kb of pathogen genome to be covered to distinguish between positive calls and background noise from the reads that came from other species.

HPV Variant Calling

RNA-seq data in FASTA format was used to identify HPV variants (Supplemental Fig. S1). Unaligned reads were taken from the PathSeq analysis (which contains HPV reads) and aligned to HPV reference genomes (HPV complete genomes from NCBI) using TopHat[48] with default parameters[49]. A BAM file containing only the HPV-related reads was generated for each sample. For each HPV isolate, a contig was generated using samtools[71] and then aligned with the HPV variant complete genome database[72] to create a phylogenetic tree using RAxML[73]. Single Nucleotide Polymorphisms (SNPs) were called from the BAM file using samtools and SNVMix[74]. The HPV variant lineages/sublineages were assigned based on the phylogenetic topology by an in-house script and confirmed visually using the SNP patterns[50].

E6 Splicing Analysis

The HPV splice junctions from RNA-seq were determined using TopHat. The splicing sites, unspliced transcripts, and their prevalence were summarized with an in-house R script that evaluated the RNA-seq reads within a window surrounding the splice sites within E6. Two transcript types were distinguished for HPV16 and HPV18: (a) transcripts that included evidence of an unspliced sequence of E6, and (b) a transcript spliced at the E6 splice donor site (position 226 for HPV16 and position and position 233 for HPV18) (spliced) (Supplemental Fig. S2). The read counts for unspliced, spliced, and the sum of both transcript types, as well as the ratio of unspliced/spliced transcripts were categorized into quartiles separately for HPV16 and HPV18 (Supplemental Table 3).

Identification of HPV Integration from RNA-seq Data (BC)

In order to assess potential genomic integration of HPV in 178 RNA-seq tumor libraries, ABySS v1.3.4[75] was used to generate *de novo* assemblies for each library, using only the reads classified by BBT (above) as human, HPV, multi-match, or no-match (Supplemental Fig. S3a). In order to address how variations in transcript abundance influence assembly[76], we generated sets of assemblies using every second k-mer length between 24 and 48 bp, and then generated a working contig set for each library by merging the contigs from all of its k-mer assemblies using Trans-ABySS v1.4.8[76]. We reran BBT on the working contig set, applying only human and HPV filters and identifying contigs that matched both filters. We identified viral-host chimeric contigs that suggested splicing of HPV donor splice sites into host splice acceptor sequences by using BLAT v34[77] to align each contig to the GRCh37-lite human reference genome and to 293 HPV reference genomes. After removing any human/viral contig that had a gap longer than 10 bp between the human- and viral-aligned segments, we retained the highest-scoring human-viral contig alignment combination. We required a contig's aligned sequences to span at least 90% of its overall length, and to overlap by less than 50%. We required a viral-human contig junction to have at least 5 mate flanking reads or 3 mate spanning reads (Supplemental Fig. S4a, b). Human splice junction contig coordinates were annotated against RefSeq and UCSC gene annotations (last modified on June 30, 2013) from the UCSC genome browser[78].

Since the chimeric contig junctions represent splicing between a viral transcript and a human transcript, the junction coordinate in each genome may not correspond to the actual location of the DNA integration, and a given genomic (i.e. DNA) integration event can be reported in RNA-seq data as multiple transcript splice sites whose genomic locations span large distances[79].

Identification of HPV Integration from RNA-seq Data (Broad Institute)

An HPV-positive sample was considered integration positive if there were at least 5 flanking reads and 10 total spanning reads (summing mate and single) supporting an integration site. Flanking read pairs were defined as having one end of the paired-end read mapped to the HPV genome and its mate pair mapped to the human genome. Spanning reads were defined as having one end of the paired-end read spanning the integration junction and its mate pair mapped to either the human or HPV genome. Once HPV reads were obtained (Pathseq, above), we extracted all pair mates and used Tophat-2.0.8[80] with fusion option enabled to map these paired end reads to a combined database containing the human genome and an HPV genome. Next, spanning reads and flanking reads were identified from the aligned BAM file. Human genes involved in the integration were identified using the breakpoint coordinates against RefSeq and UCSC gene annotations (last modified on June 30, 2013) from the UCSC genome browser[78].

Inter-Center Concordance Calls for RNA-seq Integration Events

We used a two-step approach to assess concordance between RNA-seq viral-human junction locations in the GRCh37-lite human reference genome ('sites') reported by alignments of 48-bp RNA-seq reads (BI), and of contigs with a mean length of approximately 1.5 kb that were generated from these reads by *de novo* assembly (BC) (Supplemental Fig. S3b).

We first assessed mate flanking, mate spanning and single spanning read evidence for sites (Supplemental Fig. S4a). Considering distributions of supporting evidence for three types of site calls and the number of calls from the two methods as a function of evidence strength, we set thresholds for 'confident' site calls that were 5 flanking and 3 spanning read pairs for contigs from *de novo* assembly (BC), and 5 flanking and 10 total spanning for read alignments (BI) (Supplemental Fig. S4b).

Consistent with sets of chimeric viral-human transcripts being derived from a genomic integration location, we noted that sites reported by both methods tended to occur as localized clusters. Given this clustering, sites on each chromosome were combined into a smaller set of ‘events’ using a 500-kb window and locating an event at the midpoint of its supporting sites (Supplemental Fig. S4c). The events identified by assembly were then compared with those from read alignments on both the patient and event levels (Supplemental Fig. S5).

To take advantage of differences between the contig-based and read-based integration methods, *all* method-specific integration events (both confident and non-confident events) were used for concordance analysis. An integration event was labeled as ‘concordant’ between the methods when both methods reported an event within 500 kb in the same patient. For some concordant events, both methods reported a confident event (i.e. the total read support passed the center-specific read evidence thresholds noted above). For cases in which one method called a confident event but the other a non-confident event, ‘inferred confidence’ was assigned to the concordant event. An integration event was labeled as ‘discordant’ when only one center reported a confident integration event within 500 kb.

For *intra*genic RNA-seq integration events we anticipated that most of the human transcripts associated with an event will be on the same genomic strand; however, no transcript strand information is available for intergenic integration events. For both intragenic and intergenic concordant events, we reported a range of coordinates that extends from the most proximal to the most distal supporting site (Supplemental Table 3).

For the 169 HPV-positive patients, 141 patients had integration events that were confident or inferred-confident, while the remaining 28 patients had no confident integration events. Of the 141 patients, 129 had events called by both methods, two had confident events that were called only by BC, and 10 had confident events called only by BI. Of the 129 patients with events called by both methods, all events were concordant in 90 patients. These concordant events consisted of 91 that were confident and 6 that were inferred confident. In 39 of the 129 patients, there were both concordant and discordant events. These events consisted of 43 concordant/confident events, 4 concordant/inferred confident, 1 concordant and not confident, and 57 (12 BC and 45 BI) that were discordant and confident.

Integration Calls from Low-Pass WGS Data (Harvard Medical School)

A pipeline was used that took 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, an original set of all PE reads that was mapped and unmapped to the human genome was used. Two subsets of reads were generated: ends represented by human sequences and their unmapped mates. Such unmapped reads were then aligned against the specific viral genome identified in the previous step. Clusters of discordant read pairs were calculated. In order to determine the presence of a cluster, we used an empirical cutoff of 3 discordant read pairs within the same integration region. Chimeric viral-human reads were then searched to assess the precise site of a candidate integration event at nucleotide resolution. Soft-clipped reads, in which only a portion of a read had been mapped to the human genome, were filtered from the original PE dataset and were aligned by BLAT (v.34) to the virus genome.

Integration Calls from WGS Data (Washington University in St. Louis)

WGS data for 70 tumor samples were downloaded from CGHub and aligned to a custom reference consisting of human GRCh37-lite and HPV 6, 16, 18, 31, 33, 35, 39, 45, 52, 56, 58, and 59 sequences, along with Polyoma BK, Polyoma HPyV7, Hepatitis B, Merkel Cell Polyoma as well as HHV 1, 4, and

5. Bwa v0.5.9 was used with default parameters for both bwa aln and bwa samse/sampe, using bwa's built in quality-based read trimming (-q 5).

Virus and discordant reads were discovered by parsing the realigned BAM using samtools (version 0.1.18) and standard UNIX utilities. Virus reads were detected in 66 samples, and discordant reads were observed in 65 samples. Sixty-three samples with 5 or more discordant reads were analyzed with Pindel version 0.2.5a2[81] read pair (RP) module, and human-virus breakpoints were observed for 44 of these. Breakpoint position is returned as a range of positions on both human chromosome and virus, with accuracy limited by insert size to approximately ± 1000bp.

Integration Analysis with Copy Number, mRNA Expression, and Structural Variant Data

We assessed gene-level expression relative to somatic copy number and structural variant data for genes into which we had mapped viral-human junctions from RNA or DNA sequencing data, and for genes that were associated with enhancers into which we had mapped RNA or DNA junctions. We used somatic copy number from a GISTIC2.0 "all_data_by_genes.txt" file, and normalized RSEM gene-level RNA-seq data. We assessed viral strain, viral splice donor and acceptor coordinates[82], and total read evidence for viral-human splice junctions, considering read evidence separately for the two methods. From the combined RNA and DNA evidence, we generated schematic splicing diagrams involving viral and human transcripts.

Given rank lists for SCNA and for mRNA abundance for 74 genes that contained BC HPV16 RNA-seq breakpoints and 25 genes that contained HPV18 breakpoints, we generated 100 single-sided KS p-values for the observed ranks, using a tie-tolerant KS bootstrap test (ks.boot from the R 'Matching' package, v4.8-3.4, 1000 bootstraps), and sets of 74 and 25 random numbers that were uniform between 1/178 and 1, respectively, for each KS test. P-values were corrected for multiple testing using the Benjamini-Hochberg (BH) method.

S3. DNA Sequencing and Mutation Calling

Whole Exome/Genome Sequencing (WES/WGS) Read Alignment

Data were aligned to GRCh37-lite + 42 nonredundant accessioned HPV virus sequences (ftp://genome.wustl.edu/pub/reference/GRCh37-lite-+-HPV_Redux-build/) with bwa v0.5.9. Defaults were used in both bwa aln and bwa sampe (or bwa samse if appropriate) with the exception that for bwa aln four threads were utilized (-t 4) and bwa's built in quality-based read trimming (-q 5) was used. ReadGroup entries were added to resulting SAM files using gmt sam add-read-group-tag. This SAM file was converted to a BAM file using Samtools v0.1.16, name sorted (samtools sort -n), mate pairings assigned (samtools fixmate), resorted by position (samtools sort), and indexed using gmt sam index-bam.

Read Duplication Marking and Merging

Duplicate reads from the same sequencing library were merged using Picard v1.46 MergeSamFiles and duplicates were then marked per library using Picard MarkDuplicates v1.46. Lastly, each per-library BAM with duplicates marked was merged together to generate a single BAM file for the sample. For MergeSamFiles we ran with SORT_ORDER=coordinate and MERGE_SEQUENCE_DICTIONARIES=true. For both tools, ASSUME_SORTED=true and VALIDATION_STRINGENCY=SILENT were specified. All other parameters were set to defaults. Samtools flagstat was run on each BAM file generated (per-lane, per-library, and final merged).

Low-Pass WGS Sequencing Methods

Between 500 and 700 ng of each gDNA sample were sheared to approximately 250 bp fragments using Covaris E220 and then converted to a pair-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 pair-end 2 x 51bp setup. Tumor and its matching normal were usually loaded on the same flow cell. Raw data were converted to the FASTQ format and BWA alignment was used to generate bam files.

Somatic Mutation Calling

Somatic point mutations were detected using Samtools v0.1.16 (samtools pileup –cv -A -B), SomaticSniper v1.0.2 (bam-somaticsniper -F vcf -q 1 -Q 15), Strelka v1.0.10 (with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1).

Somatic indels were detected using the GATK 1.0.5336 (-T IndelGenotyperV2 --somatic --window_size 300 -et NO_ET), retaining only those which were called as somatic, Pindel v0.2.2 (-w 10; with a config file generated to pass both tumor and normal BAM files set to an insert size of 400), Strelka v1.0.10 (with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1).

Cross Center Somatic Mutation Calls, Annotation, Readcounts, and Filtering

All high-confidence somatic mutations predicted by other centers were downloaded from the TCGA DCC from the following archive:

BCGSC: bcgsc.ca_CESC.IlluminaHiSeq_DNASeq_automated.Level_2.1.0.0.tar.gz

UCSC: ucsc.edu_CESC.IlluminaGA_DNASeq_automated.Level_2.1.0.0.tar.gz (Single nucleotide somatic mutations were identified by RADIA (RNA and DNA Integrated Analysis))[83]

Broad: broad.mit.edu_CESC.IlluminaGA_DNASeq_automated.Level_2.1.6.0.tar.gz

Readcounts supporting the tumor and variant allele for all predicted somatic mutations were extracted from exome BAM pairs using bam-readcount v0.5 (https://github.com/genome/bam-readcount).

All putative variants were annotated using Gencode 19 derived from an imported MySQL instance of Ensembl 74. Mutations in RNA genes, the coding exons of transcripts with a complete open reading frame, and at the canonical splice donor or splice acceptor were retained. Intronic variants, intergenic variants, and variants in the 3'UTR, 5'UTR, 3' flanking region, and 5' flanking region were removed.

Potential false positives due to germline cross contamination were removed by filtering all germline variants from dbSNP 137 VCF files with a GMAF>0. In order to obtain a set of high confidence somatic variants, the following minimum supporting requirements were set: Minimum tumor supporting reads $\geq$ 2, minimum tumor VAF of 10%, minimum normal reference supporting reads $\geq$ 8, and maximum normal variant supporting reads $\leq$ 1.

Previously identified, recurrent false positives identified in other TCGA exome data were filtered as previously described[84] and remaining novel recurrent somatic mutations were manually curated to identify and remove further artifacts.

Identifying Significantly Mutated Genes (SMGs)

Mutations were included for the Extended set of 192 samples, with 178 being part of the Core Freeze set. Eleven samples were identified to exhibit greater than average mutations rates and were termed "hypermutants" (somatic mutations >600). These 11 samples were removed when identifying SMGs. MutSig[6] was utilized to identify SMGs within the exome sequencing data. All 3 sample subsets without "hypermutants" (Supplemental Table 4) were analyzed using an FDR cutoff of 0.1. Significant p-values and FDR values are shown in Supplemental Table 4.

Somatic Mutation and Structural Variant Validation Methods

Library Hybrid Capture

Tumor and normal Illumina libraries were enriched by performing hybrid capture using Roche Nimblegen SeqCap EZ custom capture oligos. Genomic DNA was utilized for library construction starting material when available, and Qiagen WGA amplified DNA was used when insufficient material was available. Each sample library received unique, dual molecular barcodes prior to pooling. The target regions for somatic indels and point mutations were the 100bp region surrounding the mutation site, while for RNA-seq fusion transcript validation the flanking region of the largest introns flanking each novel exon-exon junction were targeted. Probes designed with >5 mismatches were discarded. Additional 120-mer IDT probes targeting cancer-related viruses were combined with SeqCap custom probes prior to capture. Target and probe bed files are available at http://genome.wustl.edu/pub/custom_capture/. Each sample was pooled into one of ten sets, each containing 40 or 41 samples. Each set was captured independently and sequenced on one lane of Illumina HiSeq 1T with an estimated target coverage of 200-300x.

Read Alignment

Each lane or sub-lane of data was aligned to GRCh37-lite with bwa v0.5.9. Defaults were used in both bwa aln and bwa sampe (or bwa samse if appropriate) with the exception that for bwa aln four threads were utilized (-t 4) and bwa's built-in quality-based read trimming (-q 5) was used. ReadGroup entries were added to resulting SAM files using gmt sam add-read-group-tag. This SAM file was then converted to a BAM file using Samtools v0.1.16, name sorted (samtools sort -n), mate pairings assigned (samtools fixmate), resorted by position (samtools sort), and indexed using gmt sam index-bam.

Read Duplication Marking and Merging

Reads from multiple lanes but the same sequencing library were merged, if necessary, using Picard v1.46 MergeSamFiles and duplicates were then marked per library using Picard MarkDuplicates v1.46. Lastly, each per-library BAM with duplicates marked was merged together to generate a single BAM file for the sample. For MergeSamFiles, we ran with SORT_ORDER=coordinate and MERGE_SEQUENCE_DICTIONARIES=true. For both tools, ASSUME_SORTED=true and VALIDATION_STRINGENCY=SILENT were specified. All other parameters were set to defaults. Samtools flagstat was run on each BAM file generated (per-lane, per-library, and final merged).

Somatic Variant Calling

SNV Callers

Somatic SNVs were detected using Samtools1 v0.1.16 (samtools pileup –cv -A -B), SomaticSniper2 v1.0.4 (bam-somaticsniper -F vcf -G -L -q 1 -Q 15), Strelka3 v0.4.6.2 (with default parameters except for setting isSkipDepthFilters = 1), and VarScan4 v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1).

SNV Caller Combination and Filtering

First, Samtools calls were retained if they met all of the following rules inspired by MAQ: Site is greater than 10 bp from a predicted indel of quality 50 or greater, the maximum mapping quality at the site is ≥ 40, fewer than 3 SNV calls in a 10 bp window were around the site, site is covered by at least 3 reads and less than 1,000,000,000 reads, consensus quality ≥ 20, and SNP quality ≥ 20.

After these filters were applied, Samtools and SomaticSniper calls were unioned using joinx v1.9 (https://github.com/genome/joinx; joinx sort --stable --unique). The resulting merged set of variants were additionally filtered to remove likely false positives. Bam-readcount v0.4 (https://github.com/genome/bam-readcount) was used with a minimum base quality of 15 (-b 15) to generate metrics and retained sites based on the following requirements: Minimum variant base frequency at the site of 5%, percent of reads supporting the variant on the plus strand ≥ 1% and ≤ 99% (variants failing these criteria were filtered only if the reads supporting the reference did not show a similar bias), minimum variant base count of 4, variant falls within the middle 90% of the aligned portion of the read, maximum difference between the quality sum of mismatching bases in reads supporting the variant and reads supporting the reference of 50, maximum mapping quality difference between reads supporting the variant and reads supporting the reference of 30, maximum difference in aligned read length between reads supporting the variant base and reads supporting the reference base of 25, minimum average distance to the effective 3' ends of the read for variant supporting reads of 20% of the sequenced read length, and maximum length of a flanking homopolymer run of the variant base of 5.

After this filtering, the SomaticSniper/Samtools calls were additionally filtered to high confidence variants by retaining only those sites where the average mapping quality of reads supporting the variant allele was ≥ 40 and the SomaticScore of the call was ≥ 40.

VarScan calls were retained if VarScan reported a somatic p-value ≤ 0.07, a normal frequency ≤ 5%, a tumor frequency ≥ 10%, and ≥ 2 reads supporting the variant. VarScan variants passing these criteria were then filtered for likely false positives using bam-readcount v0.4 and identical criteria as described above for SomaticSniper. Fully filtered calls as described above for SomaticSniper and VarScan were then merged with calls from Strelka using joinx v1.9 (joinx sort --stable --unique) to generate the final callset.

Indel Callers

Indels were detected using GATK5 1.0.5336 (-T IndelGenotyperV2 --somatic --window_size 300 -et NO_ET), retaining only those which were called as somatic, Pindel6 v0.2.2 (-w 10; with a config file generated to pass both tumor and normal BAM files set to an insert size of 400), Strelka3 v0.4.6.2 (with default parameters except for setting isSkipDepthFilters = 1), and VarScan4 v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1).

Indel Caller Combination and Filtering

Pindel calls were retained if they had no support in the normal data, if they had more reads reported by Pindel than reported by Samtools at the indel position, if the number of supporting reads from Pindel was ≥ 8% of the total depth at the position reported by Samtools, or if Samtools reported a depth less than 10 at the region and Pindel reported more indel supporting reads than reads mapped with gaps at the site of the call. A Fisher's Exact test p-value ≤ 0.15 was returned when comparing the number of reads with gapped alignments versus reads without in the normal vs. the tumor samples. VarScan indel calls were retained if VarScan reported a somatic p-value ≤ 0.07, a normal frequency ≤ 5%, a tumor frequency ≥ 10%, and ≥ 2 reads supporting the variant. Filtered calls from each caller as described above were merged using joinx v1.9 (joinx sort --unique --stable) to generate the final callset.

S4. Copy Number Variation (CNV) Analysis

CNV Methods

DNA processing via SNP 6.0 arrays is described in Methods. Briefly, Birdseed was used to infer a preliminary copy number at each probe locus from raw .CEL files[52]. 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[16]. 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[53]. 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 TCGA germline samples from the ovarian cancer analysis and from samples of this cohort. 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[54]. Significance of copy number alterations were assessed from the segmented data using GISTIC2.0 (Version 2.0.22)[54]. Briefly, GISTIC2.0 deconstructs somatic copy number alterations into broad and focal events and applies a probabilistic framework to identify location and significance levels of somatic copy number alterations.

Results

Somatic copy number alterations in 178 CESC tumors were determined with SNP 6.0 arrays. There were an average of 88 copy number alterations per tumor, less that ovarian and serous endometrial carcinomas but more than endometriod endometrial carcinomas[16,17]. Analysis of focal amplifications and deletions performed by the GISTIC2.0 algorithm revealed 26 focal amplifications and 37 focal deletions along with 23 whole arms that were recurrently altered. Recurrent focal amplifications were identified at 3q26.31 (*TERC, MECOM*), 3q28 (*TP63*), 7p11.2 (*EGFR*), 8q24.21 (*MYC, PVT1*), 9p24.1 (*CD274, PDCD1LG2*), 11q22.1 (*YAP1*), 13q22.1 (*KLF5*), 16p13.13 (*BCAR4*), and 17q12 (*ERBB2*). Recurrent deletions were identified at 4q35.2 (*FAT1*), 3p24.1 (*TGFBR2*), 10q23.31 (*PTEN*), and 18q21.2 (*SMAD4*). Notably, this analysis discovered novel cervical cancer driver genes, including the therapeutic targets of immune inhibitors *CD274* (PD-L1)*, PDCD1LG2* (PD-L2), and novel linc-RNA *BCAR4*. The amplifications of PDL1/2 correlated significantly ($p < 0.0001$) with cytolytic activity[18]. *BCAR4*, which has been characterized for its role in promoting metastasis, anti-estrogen resistance, and Lapatinib sensitivity in breast cancer[19], was highly amplified, fused, and greatly overexpressed compared to other tumors that do not express the gene.

Unsupervised clustering of somatic copy number alterations revealed two groups of tumors, one group with a high rate of copy number alterations and one with less ($p<0.0001$). Interestingly, these groups also showed significant clinical and molecular differences. The CN high cluster was largely composed of squamous tumors infected with HPV16 and contained significantly more tumors with *YAP1* amplifications ($p<0.0001$). The CN low cluster contained the majority of adenocarcinomas, HPV18-infected samples, and presented a novel deletion of *TGFBR2* as well as gains of *BCAR4* and *PDL1/2*.

S5. mRNA Sequencing, Analysis, and Structural Variants

RNA-seq Methods

RNA was processed as described in Methods. For further details on this processing, refer to Description file at the DCC data portal under the V2_MapSpliceRSEM workflow (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_CESC.IlluminaHiSeq_RNASeqV2.mage-tab.1.9.0/DESCRIPTION.txt).

Unsupervised Expression Clustering

Genes with >10% missing normalized RSEM values across samples were removed from the Core Freeze dataset (n=178 samples). RSEM values were then log2-transformed after first adding a constant of 1 to all values. The gene expression matrix was further filtered to only include the top 10% most variable genes by mean absolute deviation (n=1176 genes). Consensus clustering using self-organized maps was employed to identify the most robust expression clusters for between 2 to 6 clusters. Rank survey profiles for the cophenetic and silhouette widths, along with consensus cluster membership heatmaps (data not shown) suggested that a 3-cluster solution was optimal. A nearest centroid-based classifier (CLaNC) was used to identify a set of signature genes which had the lowest cross validation and prediction errors for sample membership in their respective clusters[85]. Hierarchical clustering was performed after median-centering gene expression values using Cluster 3.0[86] (uncentered correlation with centroid linkage) and visualized using JavaTreeview[87].

Identifying a Uterine Corpus Endometrial Carcinoma (UCEC) Gene Classifier

A gene expression classifier was developed to predict whether a cancer sample was from the cervix or the uterus. The data matrix of normalized gene-level RSEM values from 170 TCGA endometrial cancer samples run on the HiSeq platform was merged with the data matrix from the cervical cancer Core Freeze dataset. This merged dataset was then randomly split into a training set (87 CESC samples, 86 UCEC samples) and a test set (91 CESC samples, 84 UCEC samples). CLaNC was used to identify a set of genes in the training set which had the lowest cross-validation and prediction errors for samples being predicted as either CESC or UCEC. A t-statistic was calculated comparing each sample's expression pattern in both the training and test sets to the mean expression profile of CESC and UCEC samples in the training set to predict whether samples were CESC or UCEC. A sample was predicted to be CESC if the t-statistic vs. UCEC was significant ($p<0.05$), but was not significantly different from the CESC mean (and vice versa for the UCEC prediction). Additionally, ANOVA was used to identify differentially expressed genes (FDR<0.05) between cervical and endometrial cancers on the entire combined dataset and the expression patterns were visualized after hierarchical clustering using JavaTreeview.

Comparing CESC, UCEC, and HNSC Gene Expression Profiles

A data matrix of normalized gene-level RSEM values from 178 cervical, 170 TCGA endometrial, and 279 TCGA head and neck cancer samples run on the HiSeq platform was used to identify expression patterns across the 3 cancer types. Genes with >10% missing normalized RSEM values across samples were removed from the combined expression dataset. RSEM values were then log2-transformed after first adding a constant of 1 to all values. The gene expression matrix was further filtered to only include the top 25% most variable genes by mean absolute deviation (n=4,039 genes). Hierarchical clustering was performed after median-centering the gene expression values and the expression patterns were visualized after hierarchical clustering using JavaTreeview.

Detecting Structural Variants from RNA-seq and WGS Data

An integrative analysis was performed to identify putative driver fusions using both WGS (low-pass and hi-coverage) and RNA sequencing data. RNA-seq data for 178 cases were analyzed using the following tools:

A. *TopHat-Fusion and BreakFusion*

We ran Tophat-fusion-0.1.0 (Beta)[88] and BreakFusion-1.0.1[89] on each of the BAM files for the Core Set samples to identify fusion candidates. We further filtered the identified candidates if a) the gene fusion pairs were identified in the normal RNA libraries in the 1000 Genomes project[90]; b) the fusion breakpoints were 10 bp or more away from known splicing sites in the Refseq database; or c) they were in self-chain regions with a self-chain alignment score greater than 10.

B. *PRADA*

PRADA aligned RNA-seq reads to a composite reference database composed of whole genome and transcriptome sequences. For this analysis, we used the hg19 human genome assembly altogether with the Ensembl64 transcriptome version. Two main criteria were required to consider a gene fusion: 1) a minimum of two discordant read pairs mapping to a candidate gene pair; and 2) a minimum of one junction spanning read mapping to a junction that connected exons between the candidate gene pair, with its pair mate mapping to the either of the two genes. In order to remove false positives and artifacts, several filters were applied[91,92]. The most prominent filter was based on significant sequence similarity between the two fusion genes (using BLASTN, Expect value = 0.001), but we also filtered fusions present in a list of fusions detected in normal samples from several tissues studied by TCGA (BLCA, BRCA, HNSC, KIRC, LUAD, LUSC, and THCA) and 3 normal samples from CESC. We used SNP6 copy number data to detect whether breakpoints exist within 100 kb from the predicted junction point, which was also a relevant filter to call fusions. Also, to take into account transcript expression level, we considered fusions with transcript allele fraction (ratio of junction spanning reads to the total number of reads crossing the junction points in the reference transcripts) > 0.01[92].

C. *MapSplice*

RNA-seq data was processed and analyzed using MapSplice version 2.0.1.9 for potential gene fusions as previously described[10] to decrease the number of false positives. The resulting gene fusion list was manually curated and filtered to only include potential events where both the donor and acceptor sequences lie within known genes. To increase the confidence in the called fusions, the list of potential gene fusions was further refined to include only fusions with coverage of at least 10 reads and that had at least 2 reads bridging the breakpoint.

Detection of structural variations in low-pass WGS data (n=50) was performed using two algorithms: BreakDancer[56] and Meerkat[57]. 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 variants in the tumor and control files. The set of structural variant calls from each tumor sample was filtered by the calls from its matched normal to remove germline variants. Structural variations 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. 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 needed to fulfill the following: (1) the read identified to span the breakpoint junction hit the predicted breakpoint region uniquely by BLAT; or (2) the mate of the read spanning the breakpoint junction was mapped near the predicted breakpoint.

High-pass WGS data (n=19) were analyzed to detect somatic structural variations using two runs of BreakDancer and one run of SquareDancer (https://github.com/ding-lab/squaredancer). The predictions were unioned after filtering each set of predictions with TigraSV[93], assembly-based, and breakpoint confirmed. To detect interchromosomal breakpoints, Breakdancer 1.4.2 was run with the optional parameters "-g -h:-a -t -q 10 -d". To detect intrachomosomal breakpoints, Breakdancer 1.4.2 was run with the optional parameters "-g -h:-a -q 10 -o". Squaredancer v0.1 was run with default parameters.

Gene fusion lists generated by all methods and platforms were integrated. We identified 22 putative structural rearrangements detected by both RNA-seq and WGS (Supplemental Table 8). In total, 26 recurrent fusions were identified, of which 3 were detected by at least two RNA-seq callers (Supplemental Table 9). Furthermore, for the samples that did not have WGS data available, we extended the analysis performed on the PRADA RNA-seq fusion calls on SNP6 array copy number data to any junction points predicted by all three RNA callers described above. Among those, 74 fusions were detected by at least 2 RNA-seq callers and 60 of them showed supporting breakpoints existing within 100 kb in SNP6 array data (Supplemental Table 10).

mRNA Results

Consensus clustering was performed on RNA-seq data from 178 CESC tumor samples using 1,176 highly variable genes to identify stable subgroupings of samples. Based on this expression data, the cervical cancer samples were separated into 3 stable clusters. A gene signature was developed consisting of 300 genes which performed optimally for grouping the samples into the clusters identified by consensus clustering. Hierarchical clustering using centroid linkage resulted in the samples being grouped into 3 clusters (Supplemental Fig. S9). Functional gene annotation analysis and gene set enrichment analysis were used to identify the biologic processes involved in the separation of the cervical cancer samples into the 3 clusters. Samples in Cluster C1 contained all but 1 of the adenocarcinomas and 2 of 3 adenosquamous samples, suggesting that this is the Non-Squamous cluster. Interestingly, this cluster also includes 15 squamous cell carcinomas with expression patterns more closely related to the non-squamous cell cancers. Samples in this cluster exhibit increased expression in genes such as *EPCAM*, *CLDN3*, *ERBB4*, *RAB17*, and *KRT18*, while also showing markedly reduced expression of genes encoding several small proline-rich proteins (SPRRs), p63, and FAT2. Samples in Cluster C2 consisted entirely of squamous cell carcinomas. Genes with elevated expression in this cluster showed enrichment of ectoderm development genes and cell junction genes. Representative genes with elevated expression include 8 members of the keratin family, *ZNF750*, and *APOBEC3A*. The robust expression of keratin family member genes suggests that this cluster could be considered a Squamous Cell – Keratinizing cluster. Samples in Cluster C3 consisted entirely of squamous cell carcinomas, with the addition of 1 adenocarcinoma and 1 adenosquamous sample. Genes with elevated expression in this cluster showed enrichment of glycoprotein genes such as *EPHB2* and *TGFB2*. Samples in this cluster generally have lower expression of keratin family members, suggesting that this cluster could be considered the Squamous Cell – Non-Keratinizing cluster.

Hierarchical clustering of RNA-seq data from 75 cervical cancer cases reported in Ojesina *et al.*[8] on the 300 TCGA gene set signature resulted in 3 main clusters as in the TCGA dataset: one enriched with adenocarcinomas, one predominantly composed of squamous samples, and one exclusively composed of squamous samples (Supplemental Fig. S47). Cluster C1 contained all but 2 of the cervical adenocarcinoma cases and exhibited similar expression patterns observed in the TCGA set, namely increased expression of *EPCAM*, *CLDN3*, *ERBB4*, *RAB17* and *KRT18*. As in the TCGA set, a distinct minority of cervical squamous cell carcinomas had expression patterns more similar to those observed in adenocarcinomas. Cluster C2 consisted entirely of cervical squamous cell carcinomas and is characterized by elevated expression of genes encoding several small proline-rich proteins (SPRRs), *TP63*, *FAT2*, *KRT6A-C*, *ZNF750* and *APOBEC3A*. Like the TCGA set, this cluster could be considered a Squamous Cell-Keratinizing cluster. Cluster C3 samples contained a mixture of squamous cell carcinomas, adenocarcinomas, and adenosquamous carcinomas. As in the TCGA set, this expression cluster is characterized by elevated expression of *EPHB2* and *TGFB2*, while also exhibiting a relative decrease in keratin family gene expression when compared with samples in Cluster C2, suggesting that this cluster could be considered the Squamous Cell-Non-Keratinizing cluster. Overall, the gene expression clustering observed in the TCGA set is recapitulated in the Ojesina *et al.* data that has been previously reported.

Cervical cancer/Endometrial cancer classification: Since primary cervical cancers can be confused with endometrial cancers that involve the cervix secondarily, we developed a gene expression classifier that differentiated cervical cancers from endometrial cancers. After randomly sorting the cervical and endometrial cancer samples into a training and test set, a 14 gene classifier was identified that had the lowest prediction error in the training set, with 0 (0%) classification errors for the endometrial samples and 4 (4.4%) classification errors for the cervical samples, for an overall error rate of 2.3%. Similar results were observed when applied to the test set: 0 (0%) classification errors for the endometrial samples and 4 (4.3%) classification errors for the cervical samples, for an overall error rate of 2.3%. These 8 cervical cancer samples predicted to be endometrial cancers by expression profiling were reevaluated by study pathologists who confirmed that these samples did indeed arise from the cervix, thus we term these samples as endometrial-like (UCEC-like) cervical cancers. Interestingly, these 8 endometrial-like cervical cancers include 7 of the 9 HPV negative cancers and all but 2 of the cancers have a non-squamous cell histology. Next, gene expression profiles were compared between cervical cancers and endometrial cancers by identifying differentially expressed genes between the 2 cancer types. Unsurprisingly, the cervical and endometrial cancers tended to cluster among members of the same tissue type, except for 6 of the 8 endometrial-like cervical cancers, which clustered among the endometrial cancers. The other 2 endometrial-like cervical cancers clustered with the C1 cervical cancers, along with 1 endometrial cancer sample (Supplemental Fig. S10).

Cervical/Endometrial/Head and Neck cancer comparison: Gene expression profiles were compared between cervical (CESC), endometrial (UCEC), and head and neck cancers (HNSC). Hierarchical clustering of the different cancer samples across 4,039 highly variable genes separated the samples predominantly according to cancer type, with a few exceptions. The cervical adenocarcinomas tend to congregate in a subcluster, along with the other samples in the non-squamous expression clusters samples, and have expression patterns quite similar to those of UCEC samples. A group of about 700 genes with relatively greater expression are shared between the CESC samples in this subcluster and UCEC samples in general. Functional analysis of these genes shows overrepresentation of genes involved in embryonic morphogenesis (*HOXA9*, *HOXB2-9*) and the axoneme (6 members of the dynein family). In addition, this group of samples exhibit elevated expression of genes seen in the Non-Squamous CESC expression cluster (*ERBB4*, *RAB17*, *KRT18*) and genes highly expressed in UCECs (*ESR1* and *PGR1*). Further, a group 27 HNSC samples grouped within the CESC cluster. Interestingly, 23 of these samples are HPV-positive compared to only 13 out of 256 samples in the HNSC cluster ($p<0.0001$; Fisher's Exact test). Functional analysis of the gene expression patterns shared by HPV-positive HNSC and CESC samples may provide insights into the effects of HPV in oncogenesis. The analysis of shared genes with relatively increased expression resulted in an overrepresentation of genes involved in meiosis, including *MEI1*, *STAG3*, *SYCEP2*, and *SYCP2* which have previously been shown to be increased in HPV-positive cancers. In addition, the HPV-positive HNSC samples that group in the CESC cluster show decreased expression of a large number of genes that exhibit increased expression in the HNSC cluster. Functional analysis of these genes show overrepresentation of genes involved in ectoderm development, cell adhesion, serine-protease inhibitor activity, wound healing, and angiogenesis (Extended Data Fig. 4a).

Structural Variant Results

To characterize structural rearrangements we performed an integrative analysis of RNA-seq (n=178) and WGS data with low-pass (n=50) and deep (n=19) coverage. We identified 22 putative structural rearrangements detected by both RNA-seq and WGS (Supplemental Table 8). In total, 26 recurrent fusions were identified, of which 3 were detected by at least two RNA-seq callers (Supplemental Table 9). Examples of putative driver events are a *FGFR3-TACC3* fusion (n=1), already known in other cancer types[9,94] but not previously reported in cervical cancer, and *ZC3H7A-BCAR4* fusions (n=4). These fusions linked exon 1 of *ZC3H7A* to exon 4 of *BCAR4*. The long non-coding RNA *BCAR4* has been shown to promote estrogen-independent growth and tamoxifen resistance in breast cancer[95,96].

S6. Methylation Analysis

Sample Preparation and Hybridization

The Illumina Infinium HM450 array[58] was used to assay the Core Set of 178 TCGA cervical cancer samples. This platform includes probes for more than 480,000 CpG sites, spanning 99% of RefSeq genes. In total, 96% of CpG islands and 92% of CpG shores are represented by at least one probe. 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 an 18 µL volume. 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 12 sample BeadChips. 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 and the other with an oligomer that 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 Infinium 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

Probes having a common single nucleotide polymorphism (SNP) (defined as a SNP with a minor allele frequency > 1% as defined by the UCSC snp135common track) within 10 bp of the interrogated CpG site and probes that overlapped with a REPEAT element (as defined by RepeatMasker and Tandem Repeat Finder Masks based on UCSC hg19, Feb 2009) within 15 bp of the interrogated CpG site were identified and excluded from subsequent analyses. In addition, probes with a non-detection probability (detection p-value) greater than 0.05 in more than 25% of the samples and those associated with the Y chromosome were excluded. Probes that are mapped to multiple sites on hg19 were annotated as NA for chromosome and 0 for CpG/CpH coordinate. The final number of probes after the above exclusions was 395,552 probes.

The Illumina HumanMethylation450 array uses two different types of probes. Specifically, this array profiles the methylation status of 485,577 CpG dinucleotides, of which 72% use the Infinium type II primer extension assay where the unmethylated (red channel) and methylated (green channel) signals are measured by a single bead[58]. The remainder use the Infinium type I primer extension assay (also used in the Illumina HumanMethylation27 array) where the unmethylated and methylated signals are measured by different beads in the same color channel. Importantly, the two probe types differ in terms of CpG density, with more CpGs mapping to CpG islands for type I probes (57%) as compared with type II probes (21%). Moreover, compared with Infinium I probes, the range of beta values obtained from the Infinium II probes is smaller. In addition, the Infinium II probes have also been shown to be less sensitive for the detection of extreme methylation values and display a greater variance between replicates[97].

Clustering Analysis

Unsupervised consensus clustering was performed as implemented in the Bioconductor package ConsensusClusterPlus, with Euclidean distance and partitioning around medoids (PAM). Consensus clustering was applied to the DNA methylation data from the entire cohort, using the most variable 1% of CpG island promoter probes.

Identification of Epigenetically Silenced Genes

Epigenetically silenced genes were identified as previously described[59]. Specifically, we first identified promotor CpG sites that met several criteria: (a) at least 90% of normal samples should be clearly unmethylated ($\beta \leq 0.1$) at that site; (b) at least 5% of tumor samples should be clearly methylated ($\beta \geq 0.3$) at that site; and (c) a t-test comparing expression levels in methylated ($\beta \geq 0.3$) and unmethylated tumor samples ($\beta < 0.1$) should be significant at an FDR<0.01. A gene was defined as epigenetically silenced if at least 25% of the promoter CpG sites met all of these criteria. A total of 120 normal samples were used for this analysis, including 10 each drawn at random from the 12 TCGA projects that include normal samples, such as lung adenocarcinoma[98], breast invasive carcinoma[11], colon adenocarcinoma[99], endometrial carcinoma[17], and others. Fisher's Exact test was used to find pathways enriched with epigenetically silenced genes. Pathways with FDR<0.05 were considered significantly enriched.

HPV DNA Methylation Signatures

DNA methylation signatures derived in TCGA head and neck squamous cell carcinomas (HNSCs)[10] were applied to the Core Set of cervical tumors. The signature is represented as two sets of CpG sites at which HNSC HPV-positive samples show significantly increased or decreased methylation, respectively. Using these sets, we computed DNA hyper and hypomethylation scores as described[10].

Additionally, empirical Bayes moderated T-tests[100] were used to identify methylation differences between HPV clades A7 and A9.

Additional Analyses

Fisher's Exact test was used to test for associations of DNA methylation clusters with histology, HPV status, HPV clade, HPV integration status, EMT score, purity, APOBEC mutagenesis level, UCEC-like sample status, and the different platform cluster assignments (Extended Data Fig. 5 and Supplemental Table 13). Empirical Bayes moderated T-tests were used to identify methylation differences between groups of interest. Correlations between DNA methylation clusters and overall survival were calculated by Kaplan-Meier analysis using a log-rank test.

Results

Classifications with 2 to 7 groups were evaluated for cluster stability and fit to choose a final partition of the samples. The DNA methylation based subtypes presented here are based on a robust 3-group partition of the samples obtained using the most variable CpG island promoter features on the Illumina Infinium HM450 array (Extended Data Fig. 5). A CIMP-high (CpG island hypermethylated) cluster is characterized by widespread methylation at CpG sites within gene promoter and CpG island regions, while the CIMP-low group is distinguished by very little methylation within islands, a methylation pattern typical of healthy epithelial tissue. HPV- tumors formed a distinct cluster within the CIMP-low group with a significantly lower mean promoter methylation level than the rest of the samples in that group (t-test p-value = 0.005). The CIMP-high cluster contained most of the endocervical adenocarcinoma samples and was enriched with samples from mRNA cluster 1, miRNA cluster 4, CN-low cluster, and the Adenocarcinoma iCluster. In addition, this cluster had higher purity samples with lower EMT score as shown by boxplots in Extended Data Fig. 5b. There was no significant difference in survival between the methylation clusters (log-rank test p=0.9)

Next, we sought to capture and characterize epigenetically silenced genes. Using all Core Set 178 tumor samples and a diverse set of 120 normal samples drawn from 12 TCGA disease projects, we identified genes for which promoter methylation was normally low and where we observed increases in methylation within tumor samples that was accompanied by loss of expression, as described above. In the cervical cancer samples this procedure yielded a set of 1026 epigenetically silenced genes (Supplemental Tables 11 and 12).

The signatures of HPV16 infection derived in head and neck cancer also distinguish HPV-positive cervical tumors from HPV- tumors (Supplemental Fig. S11). Panels A and B show the distribution of DNA hyper and hypomethylation scores for head and neck and cervical cancers, respectively. Panel D shows results for HPV16 squamous cell carcinomas of the cervix, to more closely match the head and neck samples, which are all squamous cell carcinomas and predominantly of the HPV16 type.

S7. microRNA Sequencing and Analysis

Libraries and Sequencing

MicroRNA sequence (miRNA-seq) data was generated for the Core Set of 178 tumor samples using methods described previously[11]. Reads were aligned to the GRCh37/hg19 reference human genome and read count abundance was annotated against miRBase v16 stemloops and mature strands using only exact-match read alignments. Of note, BAM files that include all sequence reads are available from

CGHub (cghub.ucsc.edu)[101]. miRBase v20 was used to assign 5p and 3p mature strand (miR) names to MIMAT accession IDs.

Unsupervised Clustering

Groups of samples that had similar abundance profiles were identified using unsupervised non-negative matrix factorization (NMF) consensus clustering (v0.20.5) in R 3.1.2, with default settings[102]. The input was a reads-per-million (RPM) data matrix for the 303 (25%) most-variant 5p or 3p mature strands. After running a rank survey with 50 iterations per solution, we chose a preferred clustering solution and performed a 500-iteration run to generate the final clustering result. The preferred solution was chosen by considering profiles of the cophenetic correlation coefficient and the average silhouette width calculated from the consensus membership matrix, Kaplan-Meier survival analysis, and clinical covariate associations for a range of candidate clustering solutions. To visualize typical vs. atypical cluster members, a profile of silhouette widths was calculated from the final NMF consensus membership matrix, whereby atypical cluster members have relatively low widths.

To generate a heatmap for the NMF results, we first identified miRs that were differentially abundant between the unsupervised miRNA clusters using a SAMseq multiclass analysis (samr 2.0)[103] in R with a read-count input matrix and an FDR threshold of 0.05. For the heatmap, miRs that had the largest SAMseq scores and median abundances greater than 25 RPM were included. The RPM filtering acknowledged potential sponge effects from competitive endogeneous RNAs (ceRNAs) that can make weakly abundant miRs less influential[104,105]. Each row of the matrix was transformed by $\log_{10}$(RPM + 1) and then the pheatmap R package (v0.7.7 or v1.0.2) was used to scale and cluster only the rows, using a Euclidean distance metric and Ward clustering.

In order to show the relationship between sample order in the all-sample n=178 cohort and the squamous n=144 cohort, we used a custom Mathematica (Wolfram Research, Champaign, IL) notebook to draw a Bezier curve between each sample's position in the squamous and all-sample clustering solutions, and placed the silhouette width profiles for the two solutions on either side of the graphic for orientation.

For clinical and molecular covariates, contingency table association p-values were calculated using R, with a Chi-square or Fisher's Exact test for categorical data, and a Kruskal-Wallis test for continuous variables like EMT scores and purity.

Differentially Abundant miRs

We identified miRs that were differentially abundant between pairs of sample groups with unpaired two-class SAMseq analyses, and across sets of more than two groups with multiclass SAMseq analyses using a read-count input matrix and an FDR threshold of 0.05. For figures, filtering was done by Wilcoxon adjusted p-value > 0.05 and a median abundance less than 50 RPM in one of the two groups being compared, or across the tumor set for multiclass results. Unfiltered results are presented in Supplemental Table 14.

Relationships Between Copy Number and miRNA Abundance

In order to characterize how somatic copy number alterations (SCNA) influenced miRNA abundance, MatrixEQTL v2.1.1[106] was used to calculate Spearman correlations between a) normalized (RPM) abundance for the subset of pre-miRNAs (i.e. stemloops) that had an RPM of at least 1.0 in at least 10 of the 178 tumor samples, and b) GISTIC2 real-valued (i.e. not thresholded) SCNAs. SCNA data used Gencode v20 gene (miRNA) names, where 383 of the 476 stemloops selected by RPM above had Gencode names in the SCNA file, and another 28 had overlapping genes with SCNA records (e.g. LPP for hsa-mir-28), allowing correlations to be calculated for 411 of the RPM-selected stemloops.

Correlations were thresholded at FDR<0.05, and for a subset of the miRNAs we generated both SCNA vs RPM scatterplots and full-chromosome SCNA heatmap graphics using IGV 2.3.40. To generate a heatmap of global SCNA vs. miR-based NMF unsupervised clustering, we imported the 'seg' data and NMF clustering results into IGV v2.3.52, and ordered the samples to correspond to the 6-cluster miR-based unsupervised NMF clustering heatmap. Samples were sorted in IGV by amplification at the location of select miRNA in order to generate more focused whole-chromosome IGV graphics for a small number of miRNAs that had the strongest relationships with SCNA.

Relationships Between Methylation and miRNA Abundance

An miRNA was considered to be epigenetically controlled if BH-corrected p-values were less than 0.01 for both a) a Spearman correlation of miRNA abundance (RPM) to beta for probes in promoter regions associated with the miRNAs, and for b) a t-test of RPM between unmethylated ($\beta < 0.1$) and methylated ($\beta > 0.3$) samples (an 'epigenetically-controlled pattern').

Relationships with EMT Scores

We identified miRNAs that have been associated with EMT[62-66] and then calculated Spearman correlations between the EMT scores and RPMs for 5p and 3p mature strands for each of these miRNAs using MatrixEQTL and filtering by FDR<0.05. Heatmaps of miR abundance were generated for the miR-based unsupervised clusters for all samples (n=178) and squamous samples (n=144), sorting samples by EMT score within each unsupervised cluster and displaying only miRs whose correlations were larger than the median for each of the four cases. For *TGFBR2*, *CREBBP*, *EP300*, *SMAD4*, miR-200a, and miR-200b, we generated covariate tracks for alterations that included mutations and homozygous deletions downloaded from the cBio portal (www.cbioportal.org) and alterations in miR-200a and miR-200b (Methods and Supplemental Information S15).

miR Targeting

We assessed potential miRNA targeting for all 178 samples and then separately for the 144 squamous samples by calculating miR-mRNA and miR-protein (RPPA) Spearman correlations with MatrixEQTL v2.1.1 using gene-level normalized abundance RNA-seq (RSEM) data and normalized RPPA data. Correlations were calculated with a p-value threshold of 0.05, and then the anti-correlations were filtered at FDR<0.05. We extracted miR-gene pairs that corresponded to functional validation publications reported by miRTarBase v4.5[22]. For miR-RPPA anti-correlations, all gene names that were associated with each antibody were used. Results were displayed with Cytoscape v2.8.3.

Relationships Between Endometrial and Cervical Tumor Samples

Analyses were performed to compare miR abundance profiles between this 178-sample cervical tumor set (CESC) and the TCGA cohort of 521 uterine corpus endometrial carcinomas (UCEC). First, we generated an unsupervised clustering solution using methods described above and annotated a selected clustering solution with the CESC vs. UCEC disease type, the CESC histological types, and the UCEC-like CESC samples (see Methods and Supplemental Information S5). miRs were then identified that were differentially abundant between UCEC and CESC samples with an unpaired two-class SAMseq v2.0 analyses with FDR<0.05, as described above.

Results

NMF unsupervised consensus clustering for 178 primary tumor samples suggested a six-cluster solution (Supplemental Fig. S12a, b). Median purities varied from 0.85 to 0.59 for the clusters (Supplemental Fig. S12c). Clusters were strongly associated with histology (p=2.2e-17), HPV clade (p=0.0018), and unsupervised clusters from other molecular platforms (Supplemental Fig. S12b). miR Clusters 5 (n=30) and 6 (n=11) separated the adenocarcinoma-enriched and HPV-negative samples into two subgroups;

however, these samples were reported as a single cluster by iCluster (Adenocarcinoma cluster), PARADIGM (C2), and mRNA (C1). In contrast, the DNA methylation CIMP-high cluster was enriched only in miR Cluster 5. miRs that were differentially abundant between the clusters and also strongly abundant in at least one cluster (Supplemental Fig. S12d) included many that are known to be associated with cancer: miR-10a-5p, 21-5p, 22-3p, 143-3p, 182-5p, 203a, 205-5p, and 375. For example, for Clusters 5 and 6 noted above, both had relatively high miR-141-3p and miR-200a-3p, and relatively low miR-205-5p, while miR-10a-5p, 21-5p, 30a, and 375 were more abundant in Cluster 5 than in Cluster 6. Cluster 2 had very high levels of miR-203a, Cluster 1 had high levels of miR-143-3p and low levels of miR-200 family members, Cluster 3 had the highest levels of the oncomiR miR-21-5p, and Cluster 4 had high levels of miR-205-5p.

For the five squamous miR-based clusters (Supplemental Fig. S13), many of the same miRs were differentially and highly abundant, such as miR-21-5p, 143-3p, 203a, and 205-5p (Supplemental Fig. S13d). Four of these five clusters corresponded to clusters from the n=178 six-cluster solution (Supplemental Fig. S12 and S13f).

There were no statistically significant differences between overall survival across the miR-based clusters for n=178 (Supplemental Fig. S12e; log-rank p=0.34) or for n=144 (Supplemental Fig. S13e; p=0.13).

Differentially Abundant miRs

miRs that were differentially abundant between unsupervised clusters or other sample groups were identified by nonparametric unpaired two-class or multiclass analyses (Supplemental Table 14). miR-944[107] and 205-5p were strikingly more abundant in squamous than in adenocarcinoma samples, while miR-192-5p, 194-5p, and particularly 375 were less abundant (Supplemental Fig. S14b). Results were similar for HPV16-positive squamous vs. HPV16-positive adenocarcinoma samples (Supplemental Fig. S14c). For HPV16-positive squamous vs. HPV18-positive squamous, only miR-944 and 375 passed the FDR<0.05 threshold (Supplemental Fig. S14d). For HPV-positive vs. HPV-negative samples, miR-944 and the weakly abundant miR-767-5p and miR-105-5p were most strongly differential (Supplemental Fig. S14e).

miRs Associated With Somatic Copy Number Alterations

While somatic copy number alterations were widespread, they were relatively weakly associated with miR clusters (Supplemental Fig. S15a). Of the miRNA stem-loops whose normalized RPM abundance was cis-correlated with SCNA, those with Spearman cis-correlations of at least 0.3 had low FDRs, and scatterplots were consistent with SCNA influencing miRNA abundance (Supplemental Fig. S15b, c, d). These miRNAs included a number that were involved in potential miR-gene targeting (Supplemental Figs. S17 and S18).

Epigenetically Controlled miRNAs

The abundance of miR-10a, 17/18a/19a/20a, 141, 150, 152, and 205 appeared to be influenced by cis-DNA methylation, with miR-10a and 205 showing the clearest differences across miR-based clusters (Supplemental Fig. S16).

Functionally Validated Potential miR-gene Targeting

We assessed potential miR targeting through miR-mRNA and miR-protein (RPPA) anti-correlations for all sample and squamous only sample cohorts (FDR<0.05, (Supplemental Table 15)). Network graphics show the subset of high-confidence, FDR-thresholded anti-correlations that have been published as validated targets (Supplemental Figs. S17 and S18). The figures distinguish genes that are available only in mRNA data from those available in both mRNA and RPPA data. The figures also distinguish between anti-correlations identified with mRNA, nonphosphorylated proteins, and phosphorylated proteins. Many cancer-associated miRs were evident in the filtered anti-correlations. For example, a subnetwork involving miR-200-family miRs, the EMT-related transcription factors *ZEB1* and *ZEB2*, the Hippo effector *YAP1*, *ERBB2,* and *ERBB3* is presented in the all sample cohort. Fewer filtered targeting relationships are reported in the squamous sample cohort, some of which include *ZEB1, ZEB2*, and *ESR1*.

Comparing Endometrial and Cervical Tumors

Unsupervised NMF consensus clustering of miR abundance profiles was used to compare 521 TCGA endometrial tumor samples with the 178 cervical tumor samples. Clustering solutions appeared acceptable for between 9 and 15 clusters, which was the maximum assessed (Supplemental Fig. S19a). Details are reported for the 12-cluster solution (Supplemental Fig. S19b). In this solution, 9 clusters were exclusively or almost exclusively endometrial. Cluster 1 was almost exclusively cervical, Cluster 3 was enriched for cervical samples, with endometrial samples generally less typical cluster members, and Cluster 8 was enriched for endometrial samples. Endometrial-like cervical cancer samples were distributed across four clusters. An unpaired two-class differential abundance analysis identified miR-944 and 205-5p as far more abundant in cervical than in endometrial tumor samples (Supplemental Fig. S19c and Supplemental Table 14).

S8. Reverse Phase Protein Array (RPPA) Analysis

RPPA Experiments and Data Processing

Frozen tumors were lysed using Precellys homogenization (Cayman Chemical, Ann Arbor, Michigan) and protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mM Hepes (pH 7.4), 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 100 mM NaF, 10 mM NaPPi, 10% glycerol, 1 mM phenylmethylsulfonyl fluoride, 1 mM Na_3VO_4, and aprotinin 10 μg/mL). RPPA was performed as described previously[108]. Briefly, tumor lysate concentrations were adjusted to 1 μg/μL as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually serial-diluted in 5 two-fold dilutions with lysis buffer and printed on nitrocellulose-coated slides (Grace Bio-Labs) using an Aushon Biosystems 2470 arrayer (Billerica, MA). Slides were probed with 192 validated primary antibodies (Supplemental Table 17) followed by detection with appropriate secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG, or Rabbit anti-Goat IgG). The signal obtained was amplified using a Cytomation–catalyzed system of Avidin-Biotinylated Peroxidase (Vectastain Elite ABC kit from Vector Lab) binding to the secondary antibody and catalyzing Tyramide-Biotin (PerkinElmer) conjugation to form insoluble biotinylated phenols. Signals were visualized by a secondary streptavidin-conjugated HRP and DAB colorimetric reaction. The slides were scanned, analyzed, and quantified using Array-Pro Analyzer software (MediaCybernetics) to generate spot intensity (Level 1 data). SuperCurveGUI[109], which is available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate relative protein levels (in log2 scale). A fitted curve ("supercurve") was created with signal intensities on the Y-axis and relative log2 amounts of each protein on the X-axis using a non-parametric, monotone increasing B-spline model[108]. Raw spot intensity data were adjusted to correct spatial bias before model fitting using "control spots" arrayed across the slides[110]. A QC metric[111] was generated for each slide to determine slide quality and only slides with 0.8 on a 0-1 scale were used further. For replicate slides, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[109,112] using median-centering across antibodies (Level 3 data). Seventeen samples with low protein levels were excluded from further analysis. In total, 192 antibodies and 155 samples were analyzed. Antibodies were selected to represent the breadth of cell signaling and repair pathways[23] conditioned on a strict validation process as previously described[113]. Antibodies are labeled as "validated" and "use with caution" based on degree of validation. Raw data (Level 1), SuperCurve nonparameteric model fitting data (Level 2), and protein loading corrected data (Level 3) were deposited at the DCC.

Consensus Clustering

Consensus clustering was performed using an R package "ConsensusClusterPlus" to determine a robust number of sample clusters. Pearson correlation was used as a distance metric and Ward was used as inner and final linkage algorithm in the unsupervised hierarchical clustering analysis. Sample cluster number and membership were determined by stability evidence of 1000 resampling iterations. After consensus clustering analysis, 3 sample clusters were determined for all 155 samples.

Silhouette Clustering

The consensus clusters of 155 samples were validated by Silhouette Clustering. Euclidean distance algorithm was used to compute the pairwise dissimilarities between samples. Out of 155 samples, 115 whose Silhouette width was larger than 0.02 were retained as Silhouette Core samples for further analysis.

Heatmap Generation

The Next Generation Clustered HeatMaps (NG-CHM) tool developed at the MD Anderson Cancer Center was used to generate heatmaps for the Level 3 RPPA data. Antibody clusters were determined by unsupervised hierarchical clustering in which Pearson correlation was used as a distance metric and Ward was linkage rule. For all samples, sample clusters were supervised by the consensus clusters. For the 115 Silhouette Core samples, sample clustering employed unsupervised hierarchical clustering using Pearson correlation as a distance metric and Ward as linkage rule.

Statistical Analysis

Pathway scores were generated as described previously[7] and the differences in pathway scores between RPPA clusters were evaluated by the non-parametric Kruskal-Wallis one-way ANOVA method. Correlation between RPPA clusters and other categorical variables were detected by Chi-Squared test, while correlations with continuous variables were examined using the non-parametric Kruskal-Wallis test. The significance of survival distributions between RPPA clusters was estimated by log-rank test and visualized with Kaplan-Meier survival curves. All statistical analyses were done using R (version 3.0.2).

S9. iCLUSTER Analysis

Data

Datasets used and transformations performed are described in Methods.

iCLUSTER Method

Integrative clustering of RNA-seq, methylation, CNV, and mature-strand miRNA data was performed using R package "iCluster"[20]. The method utilizes joint latent variable model within a likelihood framework with a lasso (L1) penalty in order to select the important features creating sparse solution. The tumor subtypes are modeled as unobserved latent variables which are simultaneously estimated from the multiple data types. Expectation Maximization (EM) algorithm is implemented for maximizing the penalized log-likelihood. Using the algorithm, posterior mean of the latent factor conditional to the data is estimated and then standard k-means clustering algorithm is used to draw inference on the cluster membership of the samples. Analyses were completed with all samples and then separately by histology (squamous and adenocarcinoma).

Optimum number of clusters k together with optimum sparseness parameter λ for L1 penalty is determined using the Proportion of Deviance (POD) method where the POD can be interpreted as the sum of the absolute differences between obtained cluster block structure and theoretical (perfect) block structure. Smaller POD indicates stronger cluster distinguishability.

In order to select an adequate number of features for the iCluster, analyses were carried out using 500, 250, 100, and 50 most variable features from each dataset. There was a high degree of concordance among the resulting clustering assignments as measured by adjusted Rand Index. The results presented here are based on the 500 most variable features from each dataset.

Association analysis of clinical features and mutations with iCluster grouping was performed using Kruskal Wallis, Wilcoxon Rank-Sum, or Fisher's Exact tests. Differences in the survival of the subjects across the cluster groups were assessed using Kaplan-Meier analysis followed by a Log-Rank test. Heatmaps were made using the heatmap function in R package "NMF."

Results

All samples: Integrative clustering was carried out using the 500 most variable features from each dataset. The integrative clustering identified three clusters consisting of 50, 86, and 42 samples. The Keratin-high cluster was entirely made up of squamous samples. The Keratin-low cluster was also enriched for squamous samples, while the Adenocarcinoma cluster contained most of the adenocarcinoma samples. Association analyses between the 14 significantly mutated genes (SMGs) identified by MutSig across the three clusters were carried out using Fisher's Exact test. *KRAS* (p=9.74e-5), *ERBB3* (2.63e-3), and *HLA-A* (2.65e-2) mutations were found to be significantly associated with clusters. *KRAS* mutations were not present in the Keratin-high cluster and *HLA-A* mutations were not present in the Adenocarcinoma cluster (Fig. 2). Further association analysis of mRNA-seq expression of these SMG genes across the 3 clusters were carried out using Kruskal Wallis test, with *NFE2L2* (4.56e-11), *TGFBR2* (4.62e-8), *ERBB3* (2.14e-7), *PIK3CA* (1.17e-4), *ARID1A* (8.74e-4), and *KRAS* (3.19e-2) expression significantly associated with clusters.

Out of 178 total samples used for clustering, 112 samples had protein expression data. Association of protein expression with cluster groups was carried out using Kruskal Wallis test, and 54 proteins were significantly differentially expressed across the three clusters. Expression of Phospho-ERK (T202/Y204) (p=3.98e-2) that maps to the SMG *MAPK1* and HER2 (p=3.38e-3) that maps to *ERBB2* were found to be significantly associated with clusters. *APOBEC3A* (p=2.90e-14), *APOBEC3C* (p=1.16e-10), *APOBEC1* (p=3.20e-11), *APOBEC3B* (p=3.72e-2), and *APOBEC3G* (p=4.46e-2) gene expression were significantly different across the clusters using Kruskall Wallis Test. In addition, HPV16A vs. HPV16 non-A variants were significantly associated with the clusters (Fisher's Exact test p-value=0. 002679).

***Squamous cell carcinoma samples*:** Integrative clustering analysis on 144 samples of squamous histology identified 2 clusters with 97 and 47 samples. Association analysis with mutations in SMGs was carried out across the two clusters, with *KRAS* mutations being significantly associated with the clusters (p=0.01). mRNA-seq expression of the SMGs was assessed across the 2 clusters using Wilcoxon test, with *PIK3CA* (6.29e-6), *NFE2L2* (7.24e-6), *HLA-B* (1.07e-3), *TGFBR2* (2.82e-3), *EP300* (5.26e-3), *MAPK1* (5.47e-3), *HLA-A* (9.47e-3) and *FBXW7* (1.11e-2) significantly associated with the clusters.

Out of 144 total squamous samples, 92 samples had protein expression data. Association of protein expression with cluster groups was carried out using Wilcoxon test. Multiple proteins involved in MAPK, RTK, and Hippo pathway signaling were associated with the squamous clusters. *APOBEC3A* (p=3.09e-11), *APOBEC3C* (p=9.43e-5), *APOBEC3B* (p=1.82e-3), *APOBEC1* (p=5.13e-3), and *APOBEC3H* (p=2.73e-2) gene expression were significantly different across the clusters using Wilcoxon Test.

***Adenocarcinoma samples*:** Integrative clustering analysis on 31 adenocarcinoma samples identified 2 clusters composed of 18 and 13 samples. Associations of gene mutations were carried out across the two clusters; however, mutations in the SMGs were not significantly associated with adenocarcinoma clusters. mRNA-seq expression of the SMGs were assessed across the 2 clusters using Wilcoxon test, with *ARID1A* expression significantly associated with clusters (p=2.76e-2).

Out of 31 total adenocarcinoma samples, 18 samples had protein expression data. Association of each protein expression with cluster groups was carried out using Wilcoxon test. Multiple proteins involved in metabolism and DNA damage repair were significantly associated with clusters. Gene expression of *APOBEC3D* (p=2.39e-4) and *APOBEC1* (p=4.94e-4) were significantly different across the clusters using Wilcoxon Test.

S10. PARADIGM Analysis

Data and Algorithm

The data and algorithm are described in Methods.

Consensus Clustering of PARADIGM Inferred Pathway Activation

Consensus clustering based on the 3877 most varying features (i.e. IPLs with variance within the highest quartile) was used to identify subtypes implicated from shared patterns of pathway inference. Consensus clustering was implemented with the ConsensusClusterPlus package in R[114]. Specifically, median-centered IPLs were used to compute the squared Euclidean distance between samples, and this metric was used as the input to the ConsensusClusterPlus algorithm. Hierarchical clustering was performed using the Ward's minimum variance method (i.e. ward inner linkage option) and 80% subsampling was performed over 1000 iterations, with the final consensus matrix clustered using average linkage. The number of clusters was selected by considering the relative change in the area under the empirical cumulative distribution function (CDF) curve as well as the average pairwise item-consensus within consensus clusters. We selected k=4 as further separation provides minimal change and decreases the within-cluster consensus. Heatmap display of the top varying IPLs was generated using the heatmap.plus package in R. Differences in overall survival (OS) between PARADIGM clusters were assessed by the log-rank test, and the chi-square test was used to evaluate associations with clinical parameters (histology and HPV clade) and single platform subtypes (mRNA, copy number, methylation, miRNA, and RPPA clusters).

Pathway biomarkers of each PARADIGM cluster (vs. all others) were identified using the t-test and Wilcoxon Rank-Sum test with Benjamini-Hochberg (BH) false discovery rate (FDR) correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and showing an absolute difference in group means > 0.05 were considered. Interconnectivity between these pathway biomarkers within the PARADIGM SuperPathway was assessed, and regulatory hubs with ≥ 10 differentially activated downstream targets were selected and displayed in a heatmap.

Pathway Biomarkers Differentiating Squamous Carcinomas and Adenocarcinomas

IPLs differentially activated between squamous carcinomas (n=144) and adenocarcinomas (n=31) were identified using the t-test and Wilcoxon Rank-Sum test with BH FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and showing an absolute difference in group means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity within the SuperPathway, such that only interconnected features via regulatory interactions were retained.

Pathway constituents of the PARADIGM SuperPathway enriched among these selected features were assessed using the EASE score with BH FDR correction, and subnetworks were constructed to identify regulatory hubs with ≥ 10 outgoing regulatory edges and visualized using Cytoscape.

Interconnected complexes and features (by any edge type) showing differential activation between squamous and adenocarcinomas within the FGFR3 network neighborhood were visualized in Cytoscape. In addition, the mRNA expression levels of FGFR1 and FGFR3 were compared using Spearman rank correlation and differences in expression of these genes in squamous vs. adenocarcinomas were visualized using box plots.

In order to illustrate the difference in p63 inferred pathway activation between squamous cell carcinomas and adenocarcinomas, a heatmap of the scaled (mean 0 and standard deviation 1) p63 PARADIGM inferred activity, scaled log2-transformed mRNA expression, and GISTIC thresholded copy number levels ordered by sample histology was constructed using the heatmap.plus package in R. The log10-transformed expression of the top two differential miRNAs between squamous vs. adenocarcinoma - miR944 and miR205 - were also scaled and included in the heatmap, and the expression of these miRNAs was compared to p63 mRNA expression levels using Pearson correlation.

Pathway Biomarkers Associated With HPV Status

IPLs differentially activated between HPV Clade A9 (n=120) vs. Clade A7 (n=45) were identified using the t-test and Wilcoxon Rank-Sum test with BH FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and that showed an absolute difference in group means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity within the SuperPathway, such that only interconnected features (at least 1 interaction of any kind) were retained. Subnetworks linked through regulatory (activation or inhibition) interactions were constructed and visualized using Cytoscape, and constituent pathways of the PARADIGM SuperPathway enriched within these subnetworks were assessed using the EASE score with BH FDR correction. This analysis was also performed restricted to the squamous histology subtype (A9: n=103, A7 n=35). A similar analysis was performed to identify pathway biomarkers distinguishing HPV negative (n=9) from HPV positive (n=169) cases.

Results

Consensus clustering using the top varying PARADIGM inferred pathway levels (IPLs) yields 4 subtypes with characteristic patterns of pathway activation (Supplemental Fig. S48). Of note, 29 of 31 adenocarcinomas are clustered together in PARADIGM C2, which also contains 7 of the 9 HPV-negative cases. In addition to associations with histology, PARADIGM subtypes also show significant associations with HPV clade as well as other single platform subtypes. Highest inferred activation of FOXA2 and XBP1-2 pathways is observed within the adenocarcinoma-enriched PARADIGM C2. Key pathway features distinguishing PARADIGM cluster C4 from non-C4 cases include highest relative inferred activities of pathways involving DNA damage, MYB, and IL-12. PARADIGM cluster C3 is associated with highest inferred FOXM1 and MYC pathway activation, while the remaining PARADIGM cluster C1 samples show highest inferred activation of HIF1A, STAT6, p53, p63, p73, ARF2, and ERK signaling.

Of the 4692 PARADIGM IPLs identified as differentially activated between adenocarcinomas and squamous cell carcinomas, 1098 are connected through regulatory interactions (activation or inhibition) (Extended Data Fig. 10). Pathway enrichment and subnetwork analysis of the interconnected differential pathway features implicates higher activation of FOXA1/ER and FOXA2 pathways in adenocarcinomas. In contrast, key distinguishing features of squamous carcinomas include higher inferred activation of p53, p63, p73, AP-1, MYC, HIF1A, and MAPK signaling. Interestingly, inferred p63 activation and to a greater extent p63 mRNA expression levels show significant correlations with the two most differentially abundant miRNAs between squamous and adenocarcinomas: miR-944 and miR-205. Also of note, FGFR3 appears to have higher inferred activity in squamous carcinoma, likely attributable to higher mRNA expression levels within this histological subtype. Paradoxically, FGFR1 mRNA levels, which show a modest but significant negative correlation with FGFR3 expression, appear higher in adenocarcinomas.

A comparison of PARADIGM inferred pathway activation between Clade A7 vs. Clade A9 HPV positive samples identifies higher inferred activation of p53 and p63 signaling and lower FOXA1 signaling in the Clade A9 infected cases. These significant differences are retained when the analysis is restricted to the squamous subtype (Fig. 5a). Consistent with expectations, inferred activation of NF-kB signaling appears lower in HPV-negative relative to HPV-positive samples. Interestingly, lower inferred activity of p53 and MAPK3 signaling is also observed.

S11. APOBEC Mutagenesis Analysis

Data Deposition

Complete output of APOBEC mutagenesis analysis used for this paper in the format of Broad Institute GDAC Firehose is in the APOBEC_CESC_res3_192.7z folder placed under controlled access at: https://tcga-data-secure.nci.nih.gov/tcgafiles/tcgajamboree/CESC/APOBEC/.

In order to navigate through data all files should stay in the same folder. The "192_genome.wustl.edu_CESC.IlluminaGA_DNASeq_curated.Level_2.1.0.0.somatic.maf_sorted_report.html" Nozzle output file provides detailed legends and annotated links to all data files. A partial set of files containing Nozzle output, graphics summaries of analysis, and the most important data files are provided in the open access APOBEC_output.zip file on the TCGA Publication Page Portal.

Methods

The exome-wide prevalence of the APOBEC mutagenesis signature and the enrichment of this signature over its presence expected for random mutagenesis were evaluated as described previously[15] with some additions (see Methods). On top of previously described output, several other parameters were calculated and annotations added that characterize the prevalence of the APOBEC mutagenesis pattern in a sample and/or that are useful for downstream analyses and comparisons. The main new parameter used in this study was the minimum estimate of the number of APOBEC-induced mutations in a sample, which is given the name "APOBEC_MutLoad_MinEstimate." Values were calculated as described in Methods and are rounded to the nearest whole number.

The complete description of data files and columns in data tables are in readme files within the analysis output APOBEC_CESC_res3_192.7z folder under controlled access and within the open access APOBEC_output.zip file. The values of "APOBEC_MutLoad_MinEstimate" and category assignments for each sample are also presented in Supplemental Table 1.

Results

Prior research has identified a stringent mutation signature tCw→tTw or tCw→tGw (mutated nucleotide is capitalized; w=A or T) characteristic of mutagenesis by a subclass of APOBEC cytidine deaminases abundant in many samples of cervical and other cancer types[8,12,13,15]. In this study, 150 out of 192 exomes displayed statistically significant ($q<0.05$) enrichment (up to 6-fold) with this signature. The signature was carried by 46% of all mutations in the dataset, approaching 70% in some exomes. Even the minimum estimate accounting for the random mutagenesis resulting in a fraction of APOBEC signature mutations indicated that up to 1500 mutations in an exome can be caused by APOBECs (Supplemental Fig. S26 and APOBEC_output.zip). APOBEC mutation load strongly correlated with the total number of mutations in a sample (Extended Data Fig. 2h), suggesting that APOBEC mutagenesis is the major source of mutations in cervical cancers. HPV infection, which has been previously linked with increased APOBEC mutagenesis in head and neck cancers[14], was also correlated with a pattern of APOBEC mutagenesis in cervical cancer samples (Supplemental Fig. S27). The cause of mutagenesis may be due to high expression of *APOBEC3* genes as a result of HPV at some point during (or before) cancer development, since transcription of APOBECs is known to be induced by factors triggering the innate immune response[115]. Indeed, expression of *APOBEC3A* showed the strongest positive correlation with mutagenesis and *APOBEC3B* showed overall high expression in cancers of the dataset (Supplemental Fig. S28). Mutagenesis could also be a consequence of DNA damage response (DDR) caused by HPV[116], resulting in increased formation of single-stranded (ss) DNA – the exclusive substrate for APOBEC cytidine deaminases. Many mutations in genes with a potential role in the initiation and/or progression of cervical cancer carried the APOBEC mutagenesis signature, with *PIK3CA* harboring the most (Extended Data Fig. 2g) similar to observations in head and neck cancers[14].

S12. EMT mRNA Score Analysis

Methods

The EMT score was computed as previously described[10,21]. Briefly, the EMT score was the value resulting from the difference between the average expression of mesenchymal (M) genes minus the average expression of epithelial (E) genes. All NA values were removed from the calculation. Two-sample t-test and ANOVA were applied to each comparison accordingly. A Cox proportional hazards model was applied to assess whether the EMT score was associated with overall survival. Kalplan-Meier plots (Log-rank test) were used to display the difference between groups (the median value of EMT score of samples).

Results

EMT scores were significantly higher in UCEC-like cancers (two sample t-test, p=0.048) (Supplemental Fig. S29b). Patients with higher EMT scores had worse overall survival (p=0.0221, log-rank test between top and bottom median EMT score patient groups) (Supplemental Fig. S29a). EMT scores were associated with the subtypes defined by different molecular platforms, including methylation CIMP (p=0.024), iCluster (0.003), miRNA (p <0.001), mRNA (p=0.003), and PARADIGM (p=0.005), which suggests the association between EMT score and global molecular alterations at different levels (Supplemental Fig. S30).

S13. Functional Epigenetic Module (FEM) Analysis

FEM Algorithm

The Functional Epigenetic Module (FEM) algorithm[39] was used to identify potentially disrupted signaling pathways between groups. FEM represents a tool for the integrative analysis of DNA methylation and gene expression data that uses protein-protein-interaction (PPI) networks[117] as the backbone for identifying subnetworks of genes that are epigenetically and functionally disregulated based on a phenotype of interest. This methodology consists of two main parts: (i) computation of edge weights for connected genes in the PPI network where the weights are a composite measure of each gene's strength of association between both gene expression and DNA methylation and the phenotype of interest, and (ii) identification of subnetworks of genes where the average weight density is significantly larger than the rest of the network.

We began by subsetting the data to consist of the set of genes (G) that overlapped between the gene expression data, DNA methylation data, and genes represented in the PPI network. We then summarized DNA methylation information at the gene level by computing the average methylation of CpG sites mapping to within 200 bp of the transcription start site (TSS200). If there were no probes mapping to within 200 bp of the transcription start site, the average methylation of CpGs mapping to within the 1st exon of the gene was computed. If there were no probes mapping to within the 1st exon of the gene, the average methylation of CpGs mapping to within 1500 bp of the TSS (TSS1500) was computed. We next calculated the test-statistics, $t_g^{(R)}$ and $t_g^{(D)}$ $g = 1, 2, ..., G$, obtained from testing the association between both gene expression and DNA methylation with the phenotype of interest for each of the G genes. A composite test-statistic for each gene t_g, $g = 1, 2, ... G$ was then computed. For genes exhibiting anti-correlation between gene expression and DNA methylation (i.e. $sign(t_g^{(R)}) \neq sign(t_g^{(D)})$), composite test-statistics were taken to be the absolute difference of the DNA methylation- and gene expression-based test-statistics (i.e. $t_g = \left|t_g^{(D)} - t_g^{(R)}\right|$); otherwise, $t_g = 0$ if $sign(t_g^{(R)}) = sign(t_g^{(D)})$. Weights between connected genes, gene g and gene h, in the PPI were taken to be the average of the composite test-statistics for those two genes (i.e. $w_{gh} = \frac{1}{2}\left(t_g + t_h\right)$). Lastly, the PPI network was scanned using a version of the spin-glass algorithm[118] to identify subnetworks where the average weight density of connected genes was significantly larger than the rest of the network. The output of the FEM methodology is a series of subnetworks whose average weight density is statistically significantly greater than would be expected by chance.

The analyses described above were carried out using the Bioconductor package 'FEM' within the R statistical programming language.

Results

In an attempt to understand the implications of HPV subtype on the underlying biology of cervical tumors, we considered several different applications of the FEM methodology to the cervical cancer data. Specifically, FEM was used to identify disrupted subnetworks between HPV clade A7 and A9 tumors and HPV-positive and -negative tumors. Identification of disrupted subnetworks between these groups was carried out using all Core Set samples (n = 178) and within squamous cell carcinomas (n = 144). In addition, we also examined disrupted subnetworks between HPV A7 and A9 adenocarcinoma tumors (n = 31). There were a total of G = 6,730 genes that overlapped between the DNA methylation data, gene expression data, and the PPI network. The total space for identifying disrupted subnetworks therefore consisted of a PPI network spanned by the 6,730 overlapping genes and the interactions between them.

Identification of disrupted subnetworks between HPV-positive and HPV-negative tumors

Although only 9 out of 178 cervical tumors were HPV-negative, our analysis revealed 13 statistically significant subnetworks ($p<0.05$) when FEM was applied to the data consisting of all cervical histological subtypes (Supplemental Fig. S31 and Supplemental Table 19: Tab S1). The size of these 13 subnetworks ranged from as small as 10 genes to 44 genes. Interestingly, 3 out of the 13 identified subnetworks were centered around genes belonging to the Fibroblast Growth Family (FGF), specifically *FGF3*, *FGF4*, and *FGFR1*. Each of these genes showed statistically significant increased promoter DNA methylation (p = 1.3e-6, 6.2e-4, and 3.8e-5, respectively) and reduced expression (p = 3.4e-9, 1.6e-11, and 1.2e-6, respectively) in HPV-positive compared with HPV-negative cervical tumors. These findings are in agreement with recent data demonstrating that HPV16 E6/E7 infection (the predominant HPV subtype in these samples) partially represses the proliferation, but not the invasive potential, of cervical cancer cells stimulated by FGF2 or FGF4[119].

Restricting analysis to only the squamous cell carcinomas (n = 144), 12 statistically significant subnetworks between HPV-positive (n = 140) and HPV-negative (n = 4) tumors were identified (Supplemental Table 19: Tab S2). Similar to the results obtained from fitting FEM using all cervical histologies, 2 out of the 14 statistically significant subnetworks were centered around FGF genes, specifically *FGF3* and *FGF4*.

To see if the disrupted subnetworks between HPV-positive and HPV-negative cervical squamous cell carcinomas were specific to cervical cancer, we next applied the FEM methodology to the HSNC dataset. In a similar manner, FEM was applied to the HNSC dataset for identifying disrupted subnetworks between HPV positive (n = 36) and HPV negative (n = 243) HNSC tumors. This analysis revealed 11 statistically significant subnetworks, which ranged in size from as small as 18 genes to as large as 62 genes (Supplemental Table 19: Tab S3). Although these 11 subnetworks were largely distinct from the 12 statistically significant subnetworks between HPV positive and HPV negative cervical squamous cell tumors, there was one common subnetwork centered around Forkhead Box A2 (*FOXA2*) (Supplemental Table 19: Tabs S2, S3). Interestingly, *FOXA2* showed significantly increased promoter methylation and decreased expression in HPV positive compared to HPV negative cases in both the HNSC tumors and squamous cell cervical tumors (Supplemental Fig. S32). It is also worth noting that many of the genes contained in the *FOXA2* subnetwork showed consistent relationships between DNA methylation/gene expression and HPV status between the HNSC and squamous cell cervical tumors. These findings may suggest a common pathway(s) by which HPV exerts its effects on tumorigenesis.

Identification of disrupted subnetworks between HPV A7 and A9 tumors

We also identified disrupted subnetworks between samples infected with HPV A7 vs. A9 clades. Applying FEM to the data consisting of all cervical histological subtypes, 8 statistically significant subnetworks were identified between HPV A7 (n = 45) and HPV A9 (n = 120) tumors (Supplemental Table 19: Tab S4). Restricting analysis to only the squamous cell cervical carcinomas (n = 136) revealed 7 statistically significant subnetworks (Supplemental Table 19: Tab S5). In the analysis restricted to non-squamous cell cervical carcinomas (n = 27), 4 statistically significant subnetworks between HPV A7 (n = 8) and HPV A9 (n = 19) tumors were identified (Supplemental Table 19: Tab S6).

S14. Immune Response Gene Analysis

Immune Response Gene Expression Analysis

The Core Set (144 squamous cell carcinomas (SCCs), 31 adenocarcinomas (ACs) and 3 adenosquamous carcinomas) samples were used in this analysis and a total of 372 genes were selected based on GO 0006954 and 0006955 annotations. The gene symbols from GO selection were merged with the mRNA-seq matrix (Supplemental Table 20).

Clustering Analysis

Consensus clustering (CC) analysis was performed based on the top 300 most variable genes filtered by median absolute deviation using the ConsensusClusterPlus package in R. The gene count numbers were log-transformed and median-centered. The agglomerative hierarchical clustering algorithm using Pearson correlation distance was performed using 80% item resampling (pItem), 100% gene resampling (pFeature), a maximum of 12 cluster counts (maxk), 1,000 resampling (reps), and a random number of seed. The total number of clusters (k) was determined by the inspection of consensus cumulative distribution function (CDF) curves shape, and the relative change in area under the CDFs curve[120].

Prognostic Cluster Analysis

An *ExpressionSet* class was designed with the TCGA mRNA-seq normalized matrix for gene expression analysis (assaydata), which included the 372 immune and inflammatory response genes and the *AnnotatedDataFrame* based on the vital status presented in Supplemental Table 1 using the Biobase package in R. The areas under the curves (AUCs) were calculated based on each gene expression and the living or deceased outcomes using the rowpAUCs function (genefilter package in R). Genes that failed to accurately predict survival ($AUC \leq 0.61$) were excluded[121]. The consensus clustering analysis was performed based on the selected genes as described above.

To analyze the association of immune cytolytic activity (CYT) score with prognostic clusters and overall survival, the geometric means of *GZMA* and *PFR1* genes in SCC samples were estimated (Supplemental

Table 21)[18].

The expression of 372 genes was compared between immune response and prognostic clusters. After the estimation of the dispersion for each gene using the "estimateTagwiseDisp" function, differentially expressed (DE) genes were identified by the exact text using edgeR package. Genes with log fold-change (logFC) > 1.0 and false discovery rate (FDR) adjusted p-value<0.05 were considered.

Gene Set Enrichment Analysis (GSEA)

GSEA was performed based on the 372 immune gene expression matrix using the GSEA software and the Molecular Signature Database (MSigDB) REACTOME-c2.cp.reactome.v4.0.symbols.gmt (http://www.broad.mit.edu/gsea/). One thousand total permutations were used, and SCC versus AC and prognostic cluster C1 versus C2 were used as phenotype labels. The gene_set profile was used as the permutation type. Cytoscape software was used to create the Enrichment map. The WEB-based GEne SeT AnaLysis Toolkit (gestalt) was used to analyze common gene pathways into each gene cluster (http://bioinfo.vanderbilt.edu/webgestalt/).

Survival Analysis

The survival analyses for immune response clusters and prognostic clusters were carried out using Kaplan-Meier curve and Cox proportional-hazards regression model in Rstudio.

Results

Consensus clustering analysis identified five immune response clusters, with most ACs (n= 29) clustering together in cluster 5. Two adenosquamous samples cluster in C5 and one in C4. Among the 372 immune response genes analyzed, 83 were differentially expressed in C5 samples versus all other samples (Supplemental Table 22). Four gene clusters were differentially expressed in C5 when compared to all other samples (Supplemental Fig. S33). Gene cluster 1 (blue) contains 9 downregulated genes in C5 (*IL1A, IL1RAP, LTB4R, S100A8, S100A9, S100A12, GPR68, SPINK5 and KRT1*). *IL1A* and *IL1RAP* are involved in IL1 signaling, and *S100A8* / *S100A9* are involved in endogenous toll-like receptor signaling. Cluster 2 (red) includes 3 downregulated genes in C5 (*CD274*, *PDCD1LG2*, *AIM2*). *CD274* and *PDCD1LG2* are involved in adaptive immune response and costimulation by CD28 family signaling. Cluster 3 (magenta) includes 4 downregulated genes in C5 (*APOL3*, *CXCL9*, *CXCL10* and *CXCL11*). *CXCL9*, *CXCL10* and *CXCL11* are involved in CXCR3-mediated signaling events. *CD274* and *PDCD1LG2* genes encode PDL1 ligands PDL1 and PDL2 protein, respectively. PDL1 is expressed in various solid tumors including squamous cell carcinomas of the lung, esophagus, and head and neck[122]. These proteins suppress T-cell effector function including the cytotoxic activity and their expression is induced by inflammatory cytokines[123]. The use of PD1 immune blockage has resulted in long-term response in a subgroup of patients with lung cancer and melanoma[124,125]. The loss of AIM2 protein (absent in melanoma 2 protein) expression has been demonstrated as a prognostic marker in colorectal cancer[126], and is associated with metastatic dissemination in melanoma and cutaneous squamous cell carcinomas[127]. dsDNA viruses are sensed by AIM2, triggering inflammosome formation and IL1B release. This mechanism is a key activator of innate and adaptive immune response[128]. A recent study demonstrated that the AIM2 inflammosome is activated by HPV16 in keratinocytes[129]. *CXCL9*, *CXCL10* and, *CXCL11* genes encode CXCR3 ligand cytokines known as angiostatic CXC chemokines[130], and are potent angiogenesis inhibitors linked to cell-mediated immunity. Cluster 4 (pale green) contains upregulated genes in C5 (*ADORA1*, *DPP4*, *NFATC4*, *CRHR1*, *TCF7*, *FCGRT* and *CCR9*). *ADORA1*, *CRHR1* and *CCR9* are involved in G protein-coupled receptor binding. Cluster 5 (yellow) has 12 upregulated genes (*GPR44*, *SIGIRR*, *XBP1*, *ELF3*, *HLA-J*, *CHRNA7*, *HDAC9*, *SKAP1*, *CCBP2*, *MNX1*, *CHST4* and *ALOX15*) that are not enriched in a common pathway.

GSEA identified four significantly enriched Reactome pathways in SCCs compared with ACs. The "immune response" pathway is enriched by the "innate immune system" node and the "adaptive immune system" and its subfamily "costimulation by the CD28 family" nodes. There are 42 genes enriched and all of them are overexpressed in SCCs compared with ACs. The *CD274*, *PDCD1LG2*, *PDCD1*, *CD80*, *CD86,* and *CTLA4* genes are in the "costimulation by the CD28 family node" and are involved in T-cell activation. Other genes involved in T-cell activation that are highly expressed in SCCs include *CD8A*, *CD28*, *GZMA*, and *PRF1*. The median CYT score is 134.3 in ACs (range from 6.4 to 591.7) and 246.4 in SCCs (range from 5.9 to 4670) (p= 0.001). Together, these results suggest that the adaptive immune response is repressed in ACs compared with SCCs. Adaptive immune response and T-cell modulation have been reported as promising therapies in human cancer[131,132]. Our data suggest that the use of immune co-stimulatory molecules may be a potential therapy for cervical ACs. Based on the clustering analysis, there is a subset of SCCs with a low immunogenic profile similar to ACs. In order to determine whether immune gene expression can select groups of patients with distinct prognosis, a prognostic clustering algorithm was developed.

Prognostic clusters: ROC analysis was used to identify groups of patients with diverse prognosis in cervical carcinomas. The ROC analysis identified 47 genes with AUC> 0.61 (Supplemental Table 23). Using this set of genes, cervical carcinomas can be clustered into two different expression subtypes (Supplemental Fig. S34a), with C2 samples having worse prognosis compared with C1 samples (C1 versus C2, HR= 2.9; p= 0.002) (Supplemental Fig. S34b).

Of the 259 differentially expressed genes between prognostic cluster C1 and C2 samples (Supplemental Table 24), 57 enriched genes were identified by GSEA (Supplemental Table 25). The main nodes enriched in C1 are "immune signaling," "TCR and downstream TCR signaling pathways," "adaptive immune system," and "costimulation by CD28 family." *PDL1/2*, *PDC1*, *CD86/CTLA4*, *CD40/CD40LG*, and *CD80/CD28* genes are overexpressed in prognostic cluster C1 tumors, indicating that costimulatory and coinhibitory receptors are modulating T-cell activity. The enriched genes in prognostic cluster C2 samples are all associated to signaling by ILs (*IL1A*, *IL1R2*, *IL6*, *IL6ST*, *TRAF6*, *RIPK2* and *MAP3K7*).

When analyzing the association between CYT score and the prognostic clusters, a significantly higher CYT score was observed in C1 samples compared with samples in C2 cluster samples (Supplemental Fig. S34c). The linear regression model demonstrated that all genes associated with T-cell immune synapses are correlated with CYT, especially *PDCD1* (r^2= 0.57), *CTLA4* (r^2= 0.57), *LAG3* (r^2= 0.57) and *CD86* (r^2= 0.45) (Supplemental Table 26).

S15. MEMo Analysis

MEMo Analysis

miRNA binary alteration calls and MEMo analysis are described in Methods.

S16. Mitochondrial DNA Analysis

Analysis Methods

Aligned BAM files from whole genome sequencing (WGS) analysis were used to extract reads aligned to mitochondria and GATK[133]. Unified Genotyper was used to call SNVs and indels. Variants detected in the tumor but not in the corresponding normal were called as somatic. Somatic events were annotated using the MITOMAP database (http://www.mitomap.org/MITOMAP). Primary tumors and blood samples showed slightly different number of mitochondria, with their medians being 59 and 80, respectively. By WGS, the coverage on the MT genome was sufficient to call somatic mutations, whereas in whole exome sequencing (WES) these regions were not selected for. However, when calling mitochondrial mutations on samples using WES data, we were able to recall 71% of variants made using WGS data.

S17. RNA Splicing Analysis

Detecting RNA Splicing Events

SpliceSeq[134] was used to analyze RNA-seq data for transcript splicing variation. SpliceSeq aligns reads to splice graphs representing all protein coding isoforms of human genes in Ensembl. Percent spliced in (PSI) values are generated for each potential splice event for all samples and all genes. The type of splice events detected include exon skip (ES), retained intron (RI), alternate donor (AD), alternate acceptor (AA), mutually exclusive exon (ME), alternate promoter (AP), and alternate terminator (AT). PSI is the ratio of normalized read counts indicating the inclusion path vs. the total covering a splice event (Supplemental Fig. S37). For further details on SpliceSeq methods, see: http://bioinformatics.mdanderson.org/main/SpliceSeqV2:Methods.

To evaluate changes in splicing patterns across CESC samples, a subset of splice events demonstrating variation across tumor samples was selected. The splice event selection criterion was: 1) Minimum average expression RPKM > 1.5; 2) PSI values for 95% of the samples; 3) occurrence in a highly expressed portion of the transcript (magnitude > .3); and 4) PSI standard deviation across samples of > .25. For genes with more than one splice event meeting this criterion, the splice event with the strongest average read coverages was selected. The resulting 219 splice events represent those with the strongest differential splicing behavior across the Core Set of samples. The full set of differential splicing event PSI values is provided in Supplemental Table 28.

Results

The PSI values of selected splice events were mean-centered across samples and used to create a hierarchical clustered heatmap of sample vs PSI (Distance Metric = Correlation; Agglomeration Method = Ward). The samples clustered into 3 clusters that were further investigated in downstream analysis (Supplemental Fig. S38). Fisher's Exact tests were performed to evaluate similarity between splicing clusters and clinical data/clusters from other platforms. Splicing cluster 2 (orange) contains the majority (24 of 31, Fisher's p-value < 0.001) of adenocarcinoma samples, and therefore overlaps with many of the other adenocarcinoma-enriched platform clusters (30 of 42 Adenocarcinoma iCluster samples, 32 of 47 mRNA C1 samples, 22 of 30 miRNA C5 samples, and 31 of 45 PARADIGM C2 samples).

Splicing clusters 1 and 3 both contain predominantly squamous samples but Cluster 3 contains a smaller subset of squamous samples that displays a strikingly different pattern of alternative splicing. In general, Cluster 3 does not have strong associations with clusters identified by the other platforms so this appears to be a unique subset of squamous samples identified by splicing analysis. The only exception is an association with PARADIGM C4 (20 of 28 members are in Splicing Cluster 3; p<0.001). Cluster 3 had no significant association with clinical annotations with the exception of vital status. Only two of the 26 patients who died are in Cluster 3 (p<0.01). A review of purity values and leukocyte fraction showed that the cluster is not characterized by a high level leukocytes or low level of purity.

Several interesting splicing events distinguished the adenocarcinoma-enriched cluster C2 and the squamous-enriched clusters C1 and C3 (Supplemental Table 29). *LIMK2* expression is associated with drug resistance in many tumor types and *LIMK2* knockdown has been shown to enhance chemotherapy effectiveness[135]. The adenocarcinoma-enriched cluster showed stronger use of exon 1 as the first exon (*LIMK2a* isoform) while the squamous clusters showed stronger use of exon 3 as the first exon (*LIMK2b* isoform – missing the first LIM domain), suggesting alternate *LIMK2* expression regulation with potential impact on LIM mediated protein-protein interactions. Erbin is an adaptor protein produced by *ERBB2IP* that contributes to the oncogenic effects of HER2 and has also been a target of novel mutation specific immunotherapy[136,137]. The *ERBB2IP* exon skip event removes the PDZ domain necessary for HER2 binding. Samples in C2 show PSI values 40% lower than C1 and C3 samples, indicating that the Erbin expressed in C2 tumor samples is less capable of interacting with HER2. CD44 is a well-studied transmembrane glycoprotein with both oncogenic and tumor suppressor properties and splice variants that have been associated with metastatic progression[138,139]. The adenocarcinoma-enriched C2 samples show reduced inclusion of the *CD44* variable exons 7-14 (generally referred to as v2-v9), which add extracellular stem structure with additional binding sites for posttranslational modifications and ligand-binding[138].

Less expected was the difference in splicing patterns that distinguish the C3 from C1 samples, as both predominantly contain squamous samples (Supplemental Table 30). The C3 cluster has a moderately better survival profile than the C1 cluster ($p<0.05$; Supplemental Fig. S39). The splice events that most strongly distinguish C3 samples from C1 samples include several cancer related genes. MAGI3 is a scaffold protein that regulates LPA to inhibit migration and invasion and cooperates with PTEN to modulate AKT kinase activity related to cell survival[140,141]. The C3 cluster includes the alternate exon 8 of *MAGI3* at a higher frequency (42% increase in PSI). Exon 8 contains an annotated domain but codes for a 25-amino acid sequence between the second WW domain and the PDZ domain that interacts with PTEN, potentially altering protein interactions. HACE1 is an E3 ubiquitin ligase that is a HER2 cooperative tumor suppressor[142]. Samples in C3 include exon 7 at increased frequency (43% PSI increase). Exon 7 contains a premature stop codon leading to a truncated or degraded protein. Interestingly, many of the top splice events that define the C3 splicing cluster involve increased inclusion of an exon that introduces a premature stop codon or an alternate termination exon leading to a shortened protein product. These splicing events include ABCC3_RI_16.2, MANBA_ES_2, DAPL1_AT_4, HACE1_ES_7, and TET2_RI_4.2.

S18. Batch Effect Analysis

Analysis Methods:

Hierarchical clustering and Principal Components Analysis (PCA) were used to assess batch effects in the CESC datasets. miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), copy number variation (GW SNP 6), and protein expression (RPPA) datasets were analyzed across all CESC samples. All of the datasets were at TCGA level 3 since that is the level on which most of the analyses presented here are based. Batch effects were assessed with respect to two variables: batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of other TCGA datasets can be found at http://bioinformatics.mdanderson.org/tcgabatcheffects.

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. Samples were clustered and then annotated with colored bars at the bottom. Each color corresponds 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. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches.

miRNA Results

Supplemental Figure S40 shows clustering and PCA plots for miRNA-seq data. miRNAs with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. The figures show small batch effects by both batch ID and TSS; however, the magnitude of batch effects wasn’t high and we did not believe that it warranted batch effects correction and subsequent potential loss of important biological and technical variation in the data.

DNA Methylation Results

Supplemental Figure S41 shows clustering and PCA plots for the Infinium DNA methylation platform. Small batch effects by batch ID and TSS were seen, but once again they were deemed small enough not to warrant batch effects correction.

RNA-seqV2 Results

Supplemental Figure S42 shows clustering and PCA plots for the RNA-seq platform. Small batch effects were seen by both batch ID and TSS, but not enough to warrant algorithmic batch effects correction.

Copy Number Variation Results

Supplemental Figure S43 shows clustering and PCA plots for the copy number variation data generated on the SNP 6 platform. Small batch effects were seen by both batch ID and TSS, but not enough to warrant algorithmic batch effects correction.

Protein Expression Results

Supplemental Figure S44 shows clustering and PCA plots for the protein expression data generated on the RPPA platform. Small batch effects were seen by both batch ID and TSS, but not enough to warrant algorithmic batch effects correction.

References

68. Cancer, I.A.f.R.o. WHO Classification of Tumours of Female Reproductive Organs Vol. 6 (International Agency for Research on Cancer 2014)
69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009)
70. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997)
71. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009)
72. Burk, R.D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232-243 (2013)
73. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006)
74. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**, 730-736 (2010)
75. Simpson, J.T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117-1123 (2009)
76. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909-912 (2010)
77. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656-664 (2002)
78. Kuhn, R.M., Haussler, D. & Kent, W.J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144-161 (2013)
79. Schwartz, S. Papillomavirus transcripts and posttranscriptional regulation. *Virology* **445**, 187-196 (2013)
80. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36-R36 (2013)

81. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009)

82. Johansson, C. & Schwartz, S. Regulation of human papillomavirus gene expression by splicing and polyadenylation. *Nat. Rev. Microbiol.* **11**, 239-251 (2013)

83. Radenbaugh, A.J. *et al.* RADIA: RNA and DNA Integrated Analysis for somatic mutation detection. *PLoS One* **9**, e111516 (2014)

84. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013)

85. Dabney, A.R. ClaNC: point-and-click software for classifying microarrays to nearest centroids. *Bioinformatics* **22**, 122-123 (2006)

86. de Hoon, M.J.L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-1454 (2004)

87. Saldanha, A.J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248 (2004)

88. Kim, D. & Salzberg, S.L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72-R72 (2011)

89. Chen, K. *et al.* BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics* **28**, 1923-1924 (2012)

90. Greger, L. *et al.* Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. *PLoS One* **9**, e104567 (2014)

91. Torres-García, W. *et al.* PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* **30**, 2224-2226 (2014)

92. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* (2014)

93. Chen, K. *et al.* TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* **24**, 310-317 (2014)

94. Parker, B.C. *et al.* The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma. *J. Clin. Invest.* **123**, 855-865 (2013)

95. Godinho, M., Meijer, D., Setyono-Han, B., Dorssers, L.C.J. & van Agthoven, T. Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells. *J. Cell. Physiol.* **226**, 1741-1749 (2011)

96. Xing, Z. *et al.* lncRNA directs cooperative epigenetic regulation downstream of chemokine signals. *Cell* **159**, 1110-1125 (2014)

97. Dedeurwaerder, S. *et al.* Evaluation of the infinium methylation 450K technology. *Epigenomics* **3**, 771-784 (2011)

98. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014)

99. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012)

100. Smyth, G.K. *Limma: linear models for microarray data*. (Springer, New York, New York, USA; 2005)

101. Wilks, C. *et al.* The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* **2014** (2014)

102. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367-367 (2010)

103. Li, J. & Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22**, 519-536 (2013)

104. Mullokandov, G. *et al.* High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods* **9**, 840-846 (2012)

105. Tay, Y., Rinn, J. & Pandolfi, P.P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344-352 (2014)

106. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012)

107. Xie, H. *et al.* Novel functions and targets of miR-944 in human cervical cancer cells. *Int. J. Cancer* **136**, E230-E241 (2015)

108. 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. *Mol. Cancer Ther.* **5**, 2512-2521 (2006)

109. Hu, J. *et al.* Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986-1994 (2007)

110. Neeley, E.S., Baggerly, K.A. & Kornblau, S.M. Surface adjustment of reverse phase protein arrays using positive control spots. *Cancer Inform.* **11**, 77-86 (2012)

111. Ju, Z. *et al.* Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics* **31**, 912-918 (2015)

112. Gonzalez-Angulo, A.M. *et al.* Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clin. Proteomics* **8**, 11-11 (2011)

113. 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. *Clin. Proteomics* **6**, 129-151 (2010)

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

115. Moris, A., Murray, S.M. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Front. Microbiol.* **5** (2014)

116. McFadden, K. & Luftig, M. Interplay between DNA tumor viruses and the host DNA damage response, in *Intrinsic Immunity*, Vol. 371. (ed. B.R. Cullen) 229-257 (Springer Berlin Heidelberg, 2013)

117. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685-D690 (2011)

118. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **74**, 016110 (2006)

119. Cheng, Y.-M., Chou, C.-Y., Hsu, Y.-C., Chen, M.-J. & Wing, L.-Y.C. The role of human papillomavirus type 16 E6/E7 oncoproteins in cervical epithelial-mesenchymal transition and carcinogenesis. *Oncol. Lett.* **3**, 667-671 (2012)

120. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91-118 (2003)

121. Lasko, T.A., Bhagwat, J.G., Zou, K.H. & Ohno-Machado, L. The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inf.* **38**, 404-415 (2005)

122. Patel, S.P. & Kurzrock, R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Mol. Cancer Ther.* **14**, 847-856 (2015)

123. Ritprajak, P. & Azuma, M. Intrinsic and extrinsic control of expression of the immunoregulatory molecule PD-L1 in epithelial cells and squamous cell carcinoma. *Oral Oncol.* **51**, 221-228 (2015)

124. Brahmer, J.R. *et al.* Safety and activity of anti–PD-L1 antibody in patients with advanced cancer. *New Engl. J. Med.* **366**, 2455-2465 (2012)

125. Topalian, S.L. *et al.* Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *New Engl. J. Med.* **366**, 2443-2454 (2012)

126. Dihlmann, S. *et al.* Lack of Absent in Melanoma 2 (AIM2) expression in tumor cells is closely associated with poor survival in colorectal cancer patients. *Int. J. Cancer* **135**, 2387-2396 (2014)

127. de Koning, H.D., van Vlijmen-Willems, I.M., Zeeuwen, P.L., Blokx, W.A & Schalkwijk, J. Absent in Melanoma 2 is predominantly present in primary melanoma and primary squamous cell carcinoma, but largely absent in metastases of both tumors. *J. Am. Acad. Dermatol.* **71**, 1012-1015 (2014)

128. Unterholzner, L. *et al.* IFI16 is an innate immune sensor for intracellular DNA. *Nat. Immunol.* **11**, 997-1004 (2010)

129. Reinholz, M. *et al.* HPV16 activates the AIM2 inflammasome in keratinocytes. *Arch. Dermatol. Res.* **305**, 723-732 (2013)

130. Strieter, R.M. *et al.* Cancer CXC chemokine networks and tumour angiogenesis. *Eur. J. Cancer* **42**, 768-778 (2006)

131. Naidoo, J., Page, D.B. & Wolchok, J.D. Immune modulation for cancer therapy. *Br. J. Cancer* **111**, 2214-2219 (2014)

132. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252-264 (2012)

133. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011)

134. Ryan, M.C., Cleland, J., Kim, R., Wong, W.C. & Weinstein, J.N. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics* **28**, 2385-2387 (2012)

135. Gamell, C., Schofield, A.V., Suryadinata, R., Sarcevic, B. & Bernard, O. LIMK2 mediates resistance to chemotherapeutic drugs in neuroblastoma cells through regulation of drug-induced cell cycle arrest. *PLoS One* **8**, e72850 (2013)

136. Tao, Y. *et al.* Role of Erbin in ErbB2-dependent breast tumor growth. *Proc. Natl. Acad. Sci. USA* **111**, E4429-E4438 (2014)

137. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **344**, 641-645 (2014)

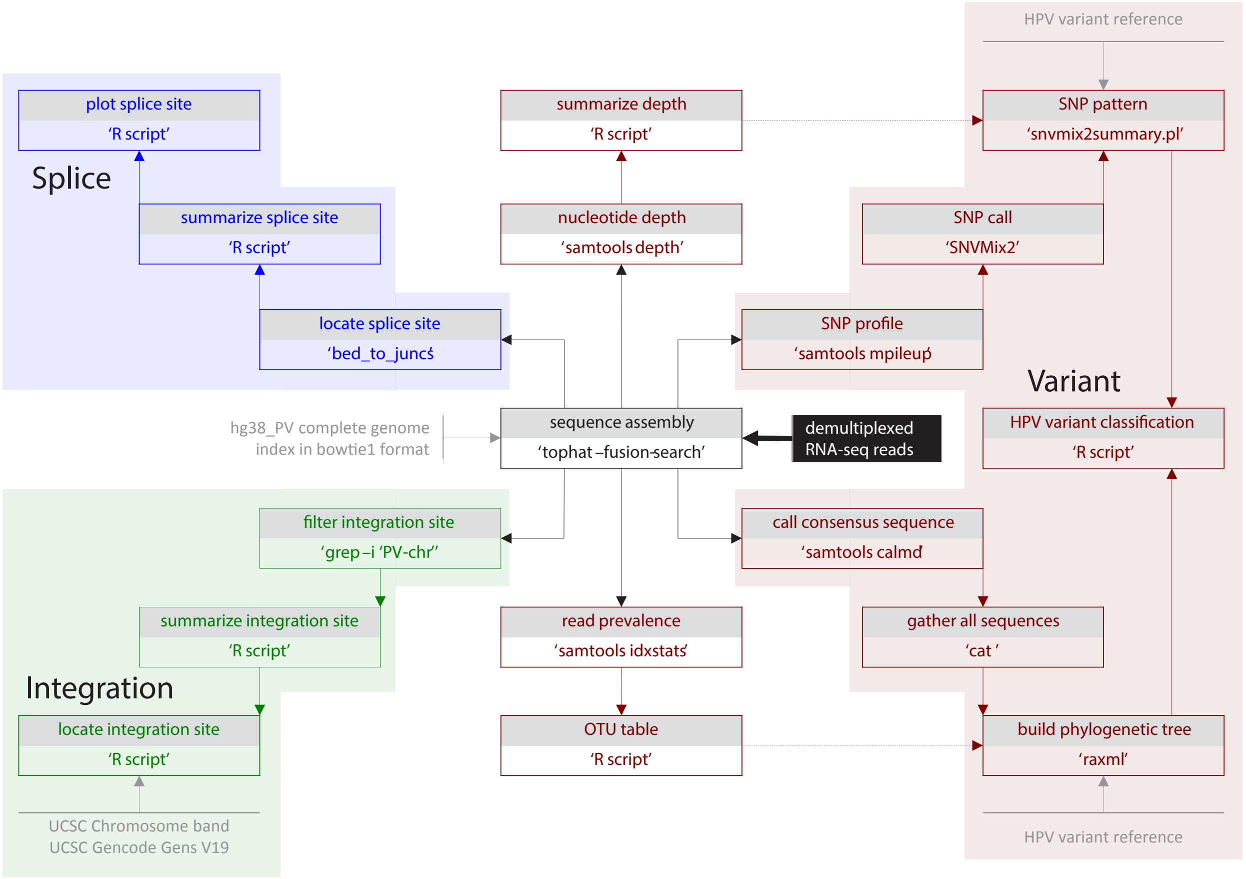
138. Louderbough, J.M.V. & Schroeder, J.A. Understanding the dual nature of CD44 in breast cancer progression. *Mol. Cancer Res.* **9**, 1573-1586 (2011)

139. Speiser, P. *et al.* CD44 is an independent prognostic factor in early-stage cervical cancer. *Int. J. Cancer* **74**, 185-188 (1997)

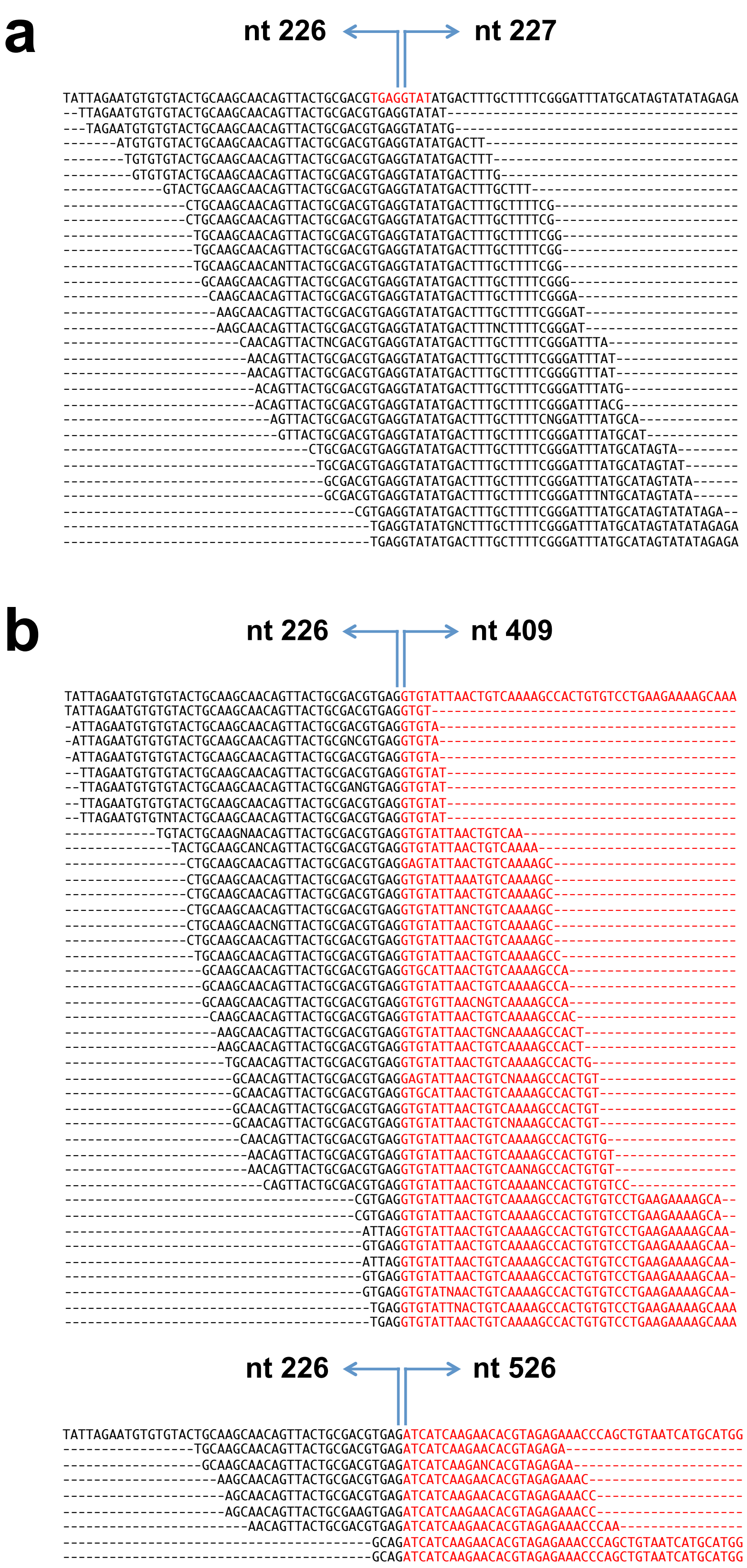
140. Lee, S.J. *et al.* MAGI-3 competes with NHERF-2 to negatively regulate LPA2 receptor signaling in colon cancer cells. *Gastroenterology* **140**, 924-934 (2011)

141. Wu, Y. *et al.* Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of MAGI3, a novel membrane-associated guanylate kinase. *J. Biol. Chem.* **275**, 21477-21485 (2000)

142. Goka, E.T. & Lippman, M.E. Loss of the E3 ubiquitin ligase HACE1 results in enhanced Rac1 signaling contributing to breast cancer progression. *Oncogene* (2015)



Supplemental Figure S1: HPV16 variant calling analysis pipeline.

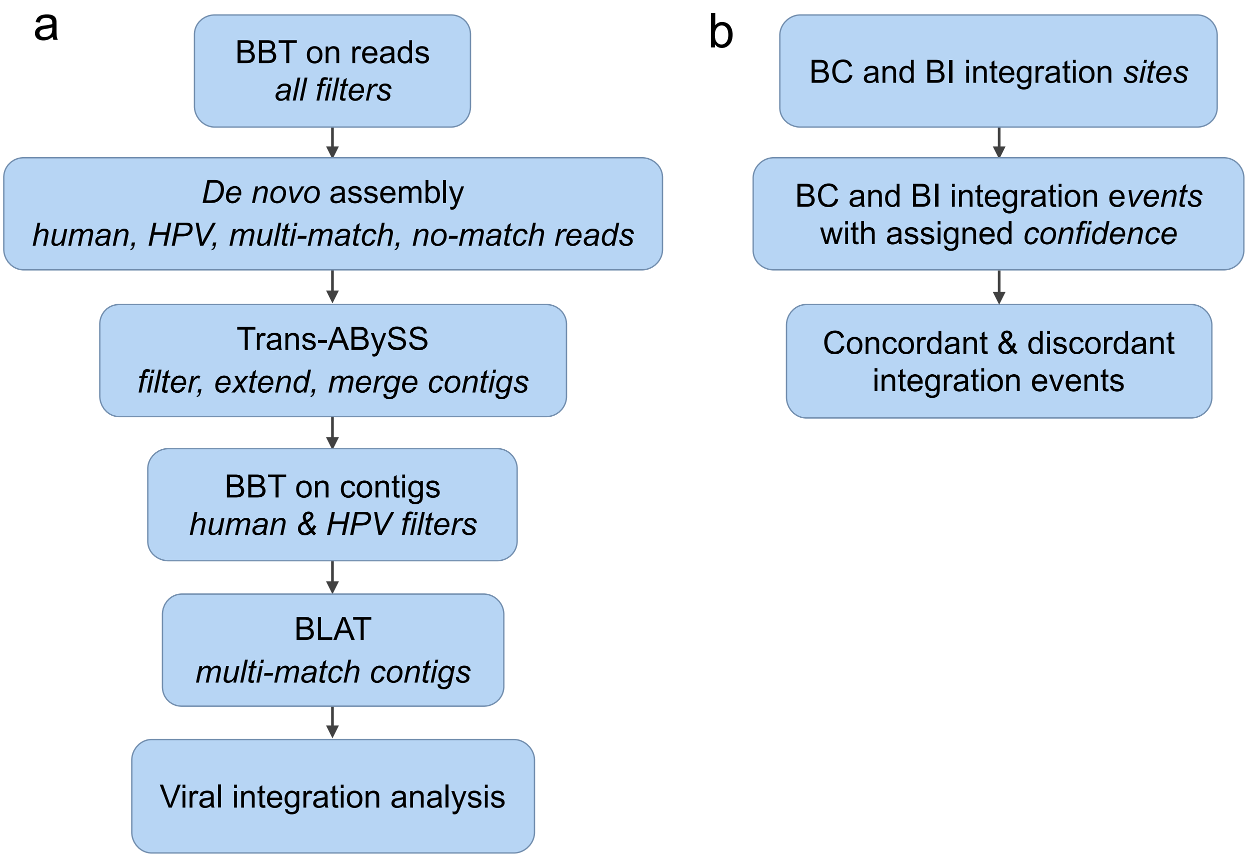


```
TATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTATATAGAGA
--TTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATAT--------------------------------------
---TAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATG-------------------------------------
-------ATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTT---------------------------------
--------TGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTT--------------------------------
---------GTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTG-------------------------------
-------------GTACTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTT---------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCG------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCG------------------------
-----------------TGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGG-----------------------
-----------------TGCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGG-----------------------
-----------------TGCAAGCAACANTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGG-----------------------
------------------GCAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGG----------------------
-------------------CAAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGA---------------------
--------------------AAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGAT--------------------
--------------------AAGCAACAGTTACTGCGACGTGAGGTATATGACTTTNCTTTTCGGGAT--------------------
-----------------------CAACAGTTACTNCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTA-----------------
------------------------AACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTAT----------------
------------------------AACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGGTTTAT----------------
-------------------------ACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATG---------------
-------------------------ACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTACG---------------
---------------------------AGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCNGGATTTATGCA-------------
----------------------------GTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCAT------------
--------------------------------CTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTA--------
---------------------------------TGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTAT-------
----------------------------------GCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTATA------
----------------------------------GCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTNTGCATAGTATA------
--------------------------------------CGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTATATAGA--
----------------------------------------TGAGGTATATGNCTTTGCTTTTCGGGATTTATGCATAGTATATAGAGA
----------------------------------------TGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTATATAGAGA
```

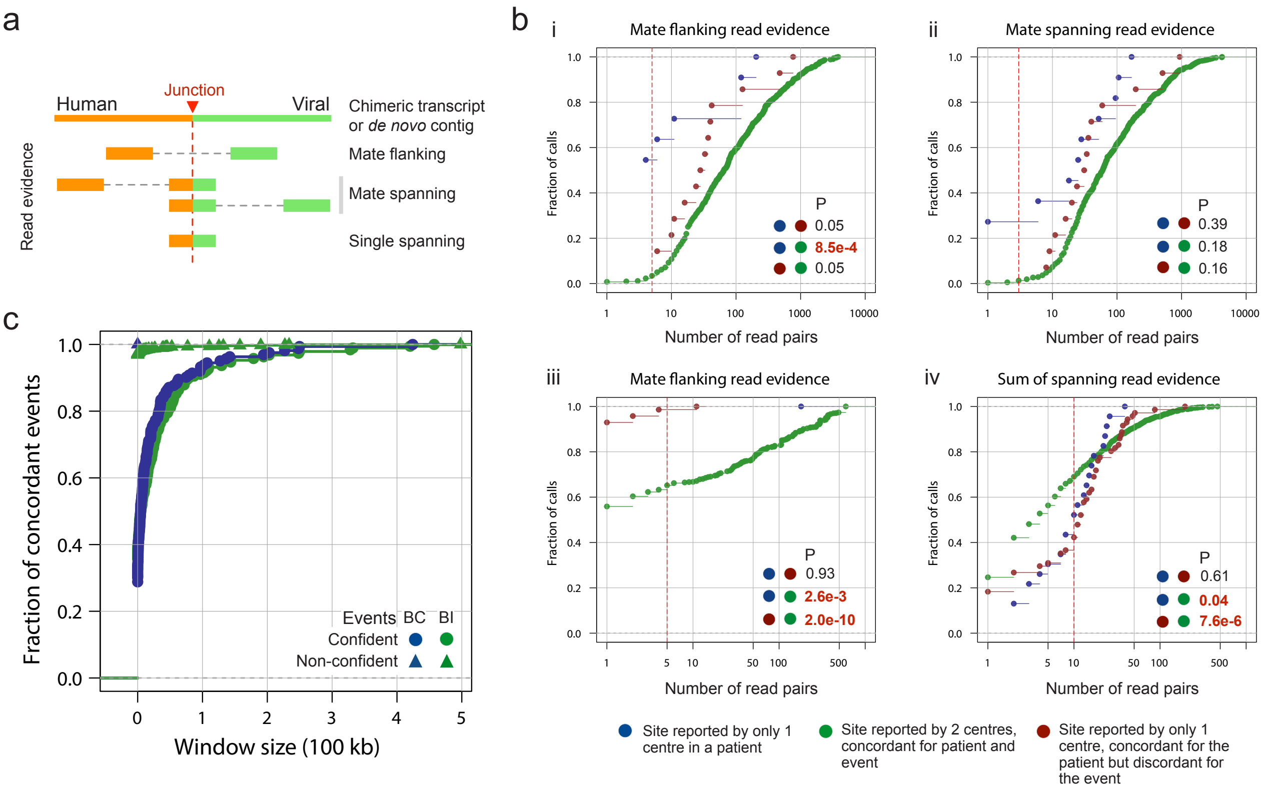
```
TATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAAA
TATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGT----------------------------------------
-ATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTA---------------------------------------
-ATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGNCGTGAGGTGTA---------------------------------------
-ATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTA---------------------------------------
--TTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTAT--------------------------------------
--TTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGANGTGAGGTGTAT--------------------------------------
--TTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTAT--------------------------------------
--TTAGAATGTGTNTACTGCAAGCAACAGTTACTGCGACGTGAGGTGTAT--------------------------------------
------------TGTACTGCAAGNAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAA----------------------------
--------------TACTGCAAGCANCAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAA--------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGAGTATTAACTGTCAAAAGC------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTGTATTAAATGTCAAAAGC------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGC------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTGTATTANCTGTCAAAAGC------------------------
----------------CTGCAAGCAACNGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGC------------------------
----------------CTGCAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGC------------------------
-----------------TGCAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCC-----------------------
------------------GCAAGCAACAGTTACTGCGACGTGAGGTGCATTAACTGTCAAAAGCCA----------------------
------------------GCAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCA----------------------
------------------GCAAGCAACAGTTACTGCGACGTGAGGTGTGTTAACNGTCAAAAGCCA----------------------
-------------------CAAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCAC---------------------
--------------------AAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGNCAAAAGCCACT--------------------
--------------------AAGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACT--------------------
---------------------TGCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACTG-------------------
----------------------GCAACAGTTACTGCGACGTGAGGAGTATTAACTGTCNAAAGCCACTGT------------------
----------------------GCAACAGTTACTGCGACGTGAGGTGCATTAACTGTCAAAAGCCACTGT------------------
----------------------GCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACTGT------------------
----------------------GCAACAGTTACTGCGACGTGAGGTGTATTAACTGTCNAAAGCCACTGT------------------
-----------------------CAACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACTGTG-----------------
------------------------AACAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAAGCCACTGTGT----------------
------------------------AACAGTTACTGCGACGTGAGGTGTATTAACTGTCAANAGCCACTGTGT----------------
--------------------------CAGTTACTGCGACGTGAGGTGTATTAACTGTCAAAANCCACTGTGTCC--------------
--------------------------------------CGTGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCA--
--------------------------------------CGTGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCA--
---------------------------------------ATTAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAA-
---------------------------------------GTGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAA-
---------------------------------------ATTAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAA-
---------------------------------------GTGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAA-
---------------------------------------GTGAGGTGTATNAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAA-
----------------------------------------TGAGGTGTATTNACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAAA
----------------------------------------TGAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAAGAAAAGCAAA
```

```
TATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGTGAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGCATGG
-----------------TGCAAGCAACAGTTACTGCGACGTGAGATCATCAAGAACACGTAGAGA-----------------------
------------------GCAAGCAACAGTTACTGCGACGTGAGATCATCAAGANCACGTAGAGAA----------------------
--------------------AAGCAACAGTTACTGCGACGTGAGATCATCAAGAACACGTAGAGAAAC--------------------
---------------------AGCAACAGTTACTGCGACGTGAGATCATCAAGAACACGTAGAGAAACC-------------------
---------------------AGCAACAGTTACTGCGAAGTGAGATCATCAAGAACACGTAGAGAAACC-------------------
------------------------AACAGTTACTGCGACGTGAGATCATCAAGAACACGTAGAGAAACCCAA----------------
----------------------------------------GCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGCATGG
----------------------------------------GCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGCATGG
```

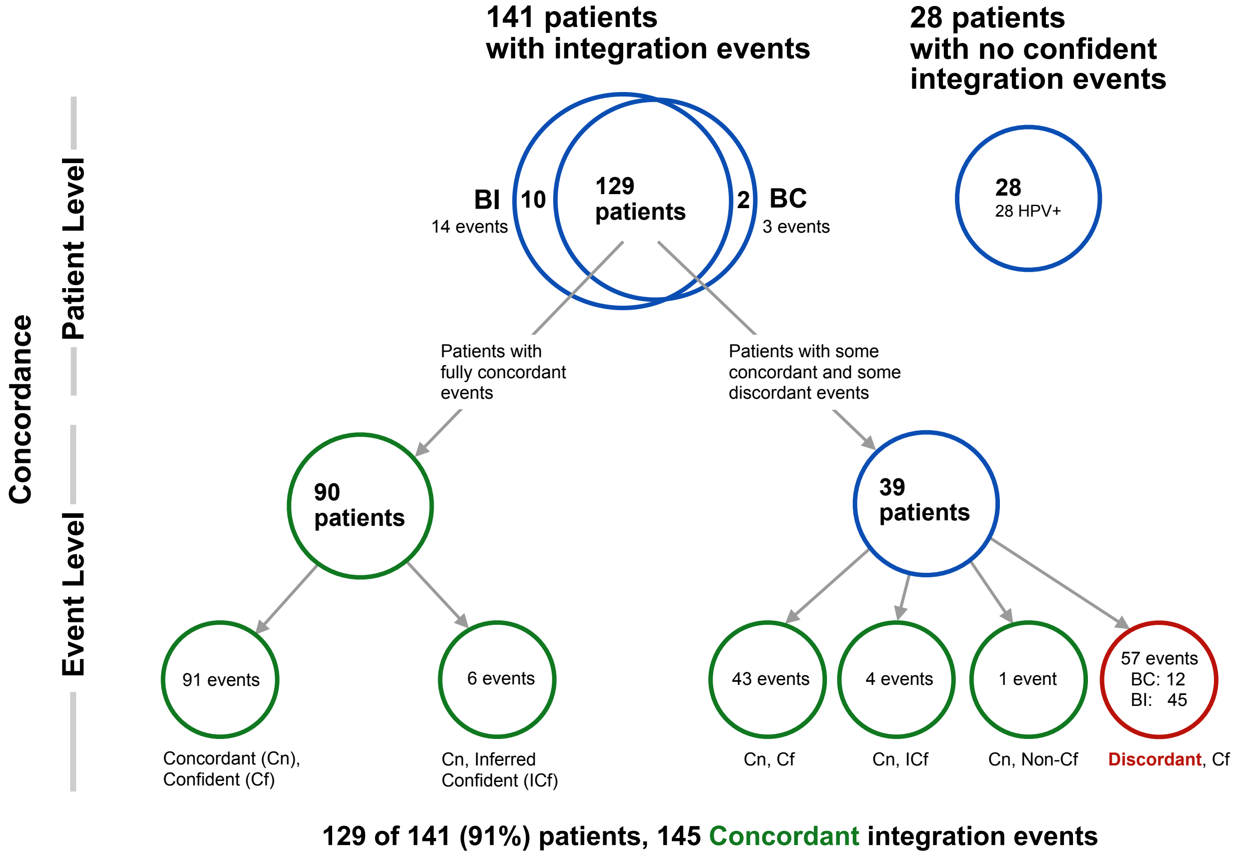
Supplemental Figure S2: Examples of HPV16 reads that indicate unspliced (a) and spliced (b) E6 transcripts.


a
BBT on reads
all filters
De novo assembly
human, HPV, multi-match, no-match reads
Trans-ABySS
filter, extend, merge contigs
BBT on contigs
human & HPV filters
BLAT
multi-match contigs
Viral integration analysis
b
BC and BI integration *sites*
BC and BI integration *events* with assigned *confidence*
Concordant & discordant integration events

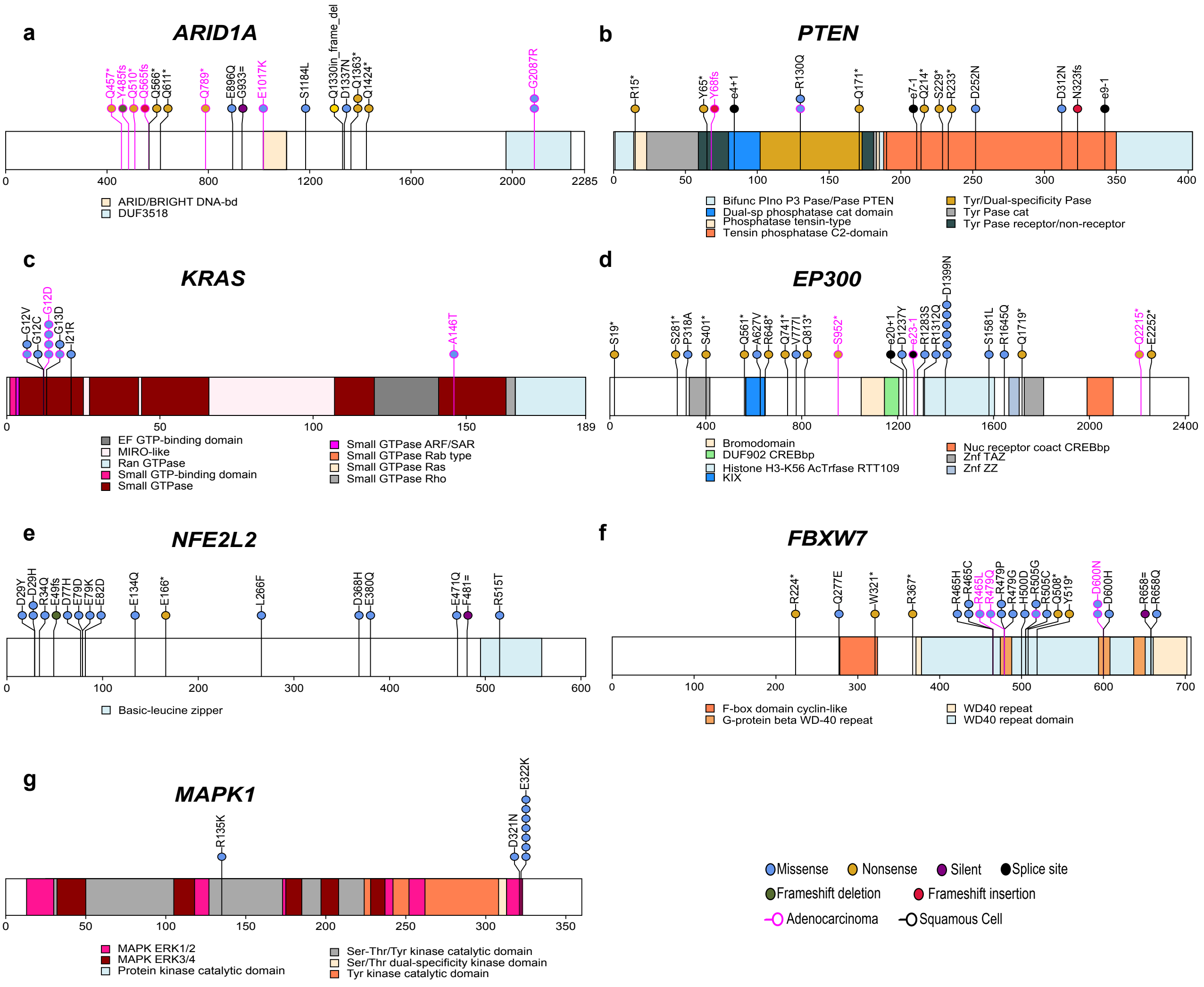
Supplemental Figure S3: Workflows for integration and concordance analyses. **a,** Viral integration workflow. **b,** Concordance analysis workflow.



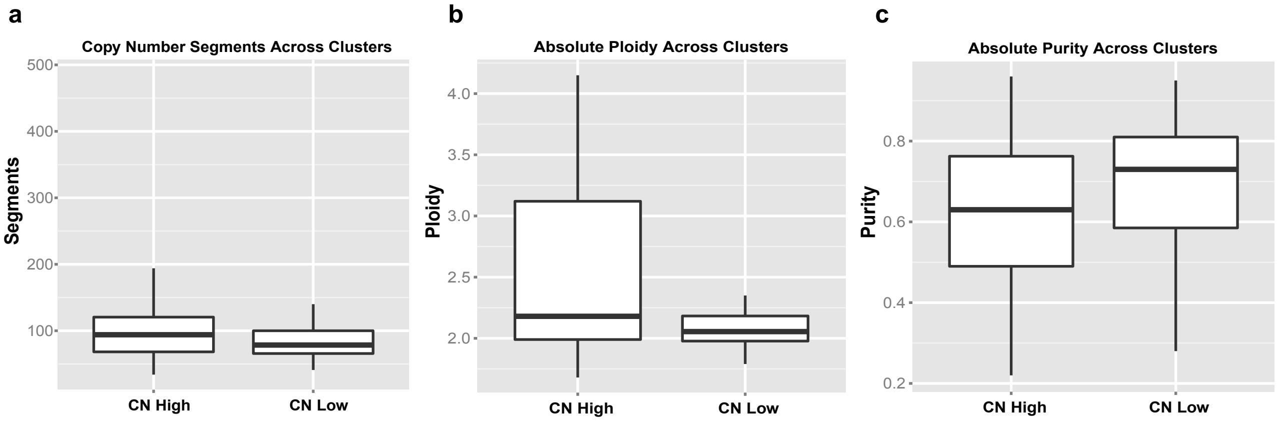
Supplemental Figure S4: Concordance analysis for integration events for RNAseq data. **a,** Types of read evidence for integration sites from RNAseq data. **b,** Distribution functions for flanking and spanning read evidence, with dashed red lines showing evidence thresholds. BC: **i)** mate flanking = 5 read pairs, **ii)** mate spanning = 3 read pairs. BI: **iii)** mate flanking = 5 read pairs, **iv)** the sum of mate spanning and single spanning = 10 read pairs. P-values are from two-sided Kolmogorov-Smirnov tests. **c,** Distribution function of the number of concordant events as a function of the window size used to group sites into events.



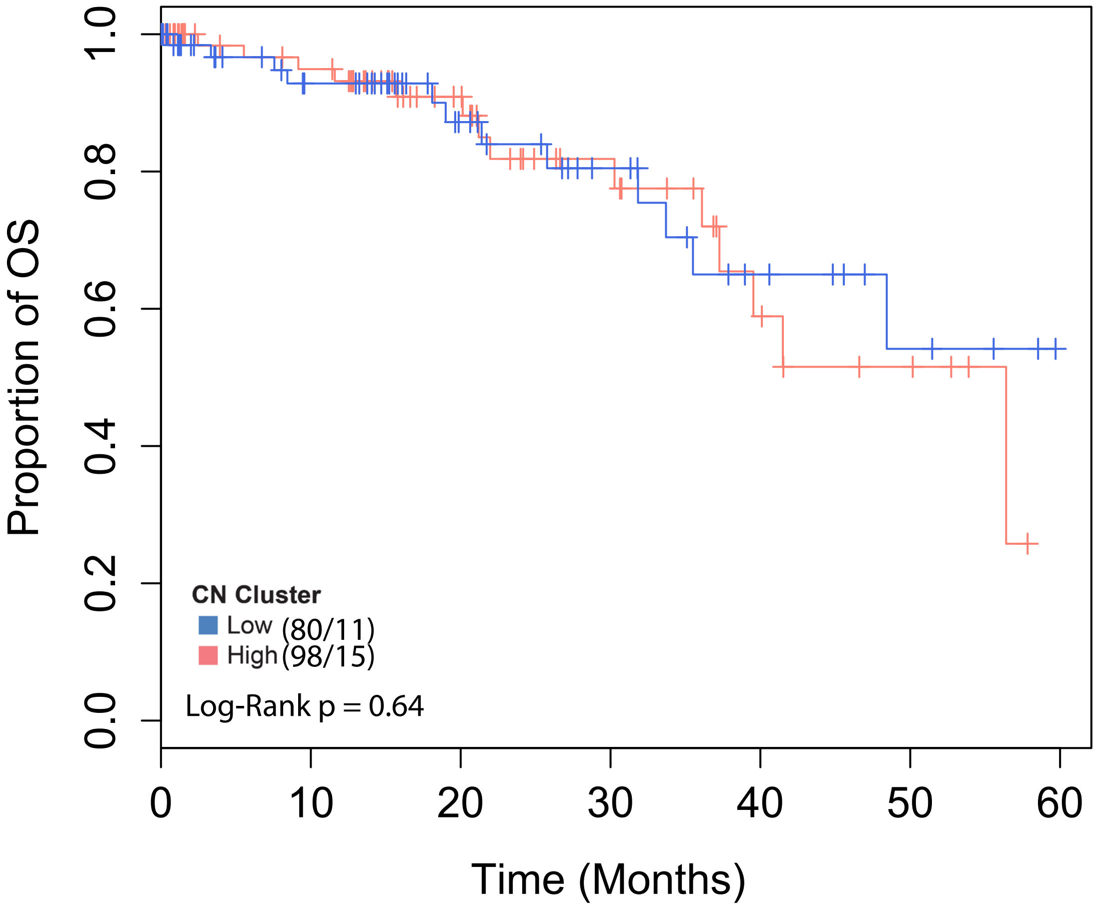
Supplemental Figure S5: Event-level concordance in 169 HPV+ patients using 500-kb windows. The upper third of the figure reports concordance at the patient level, while the lower two thirds of the figure reports concordance at the event level.



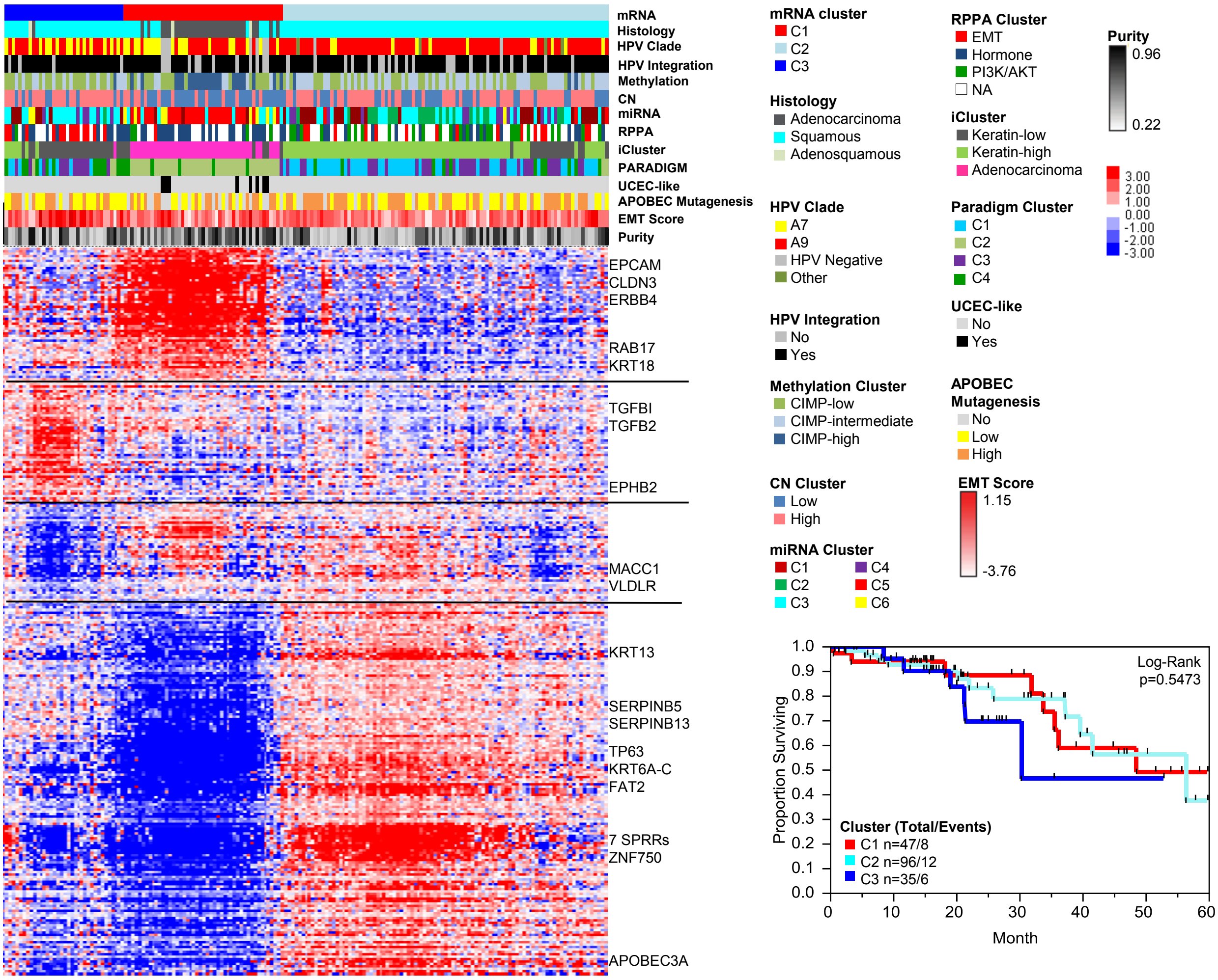
Supplemental Figure S6: Significantly Mutated Genes. **a-g,** High-confidence somatic mutations in significantly mutated genes (SMGs) among 192 exome-sequenced samples in the Extended dataset are shown. SMGs presented in Extended Data Fig. 2 are not shown here. Domains are labeled in accordance with Gencode 19 corresponding to Ensembl 74 and represent UniProt functional domains. Vertical lines indicate the boundaries of multiple annotation sources within common domain annotations as outlined in Supplemental Table 5. Mutations at canonical intronic splice donor (e+1 and e+2) and splice acceptor (e-1 and e-2) are labeled based on proximity to the nearest coding exon, ***e***. Circles represent a single mutation and are colored based on mutation type. Mutations present in squamous cell carcinomas are outlined in black while those present in adenocarcinomas are outlined in pink



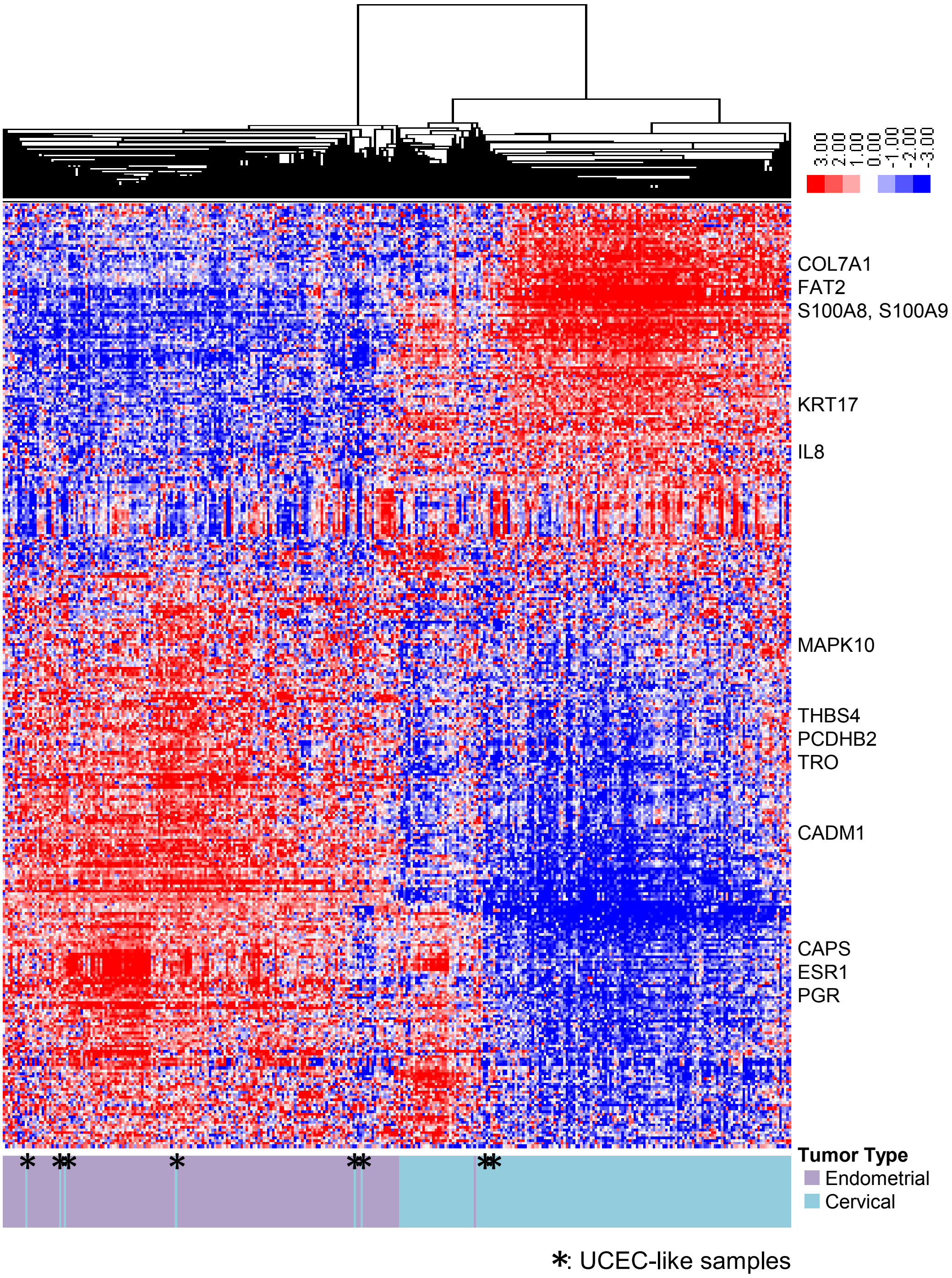
Supplemental Figure S7: Copy number segments, purity, and ploidy across CN clusters. a, Comparison of the number of copy number segments per tumor between the CN High and CN Low clusters. **b-c**, Comparison of the ABSOLUTE ploidy (**b**) and purity (**c**) per tumor between the CN High and CN Low clusters.



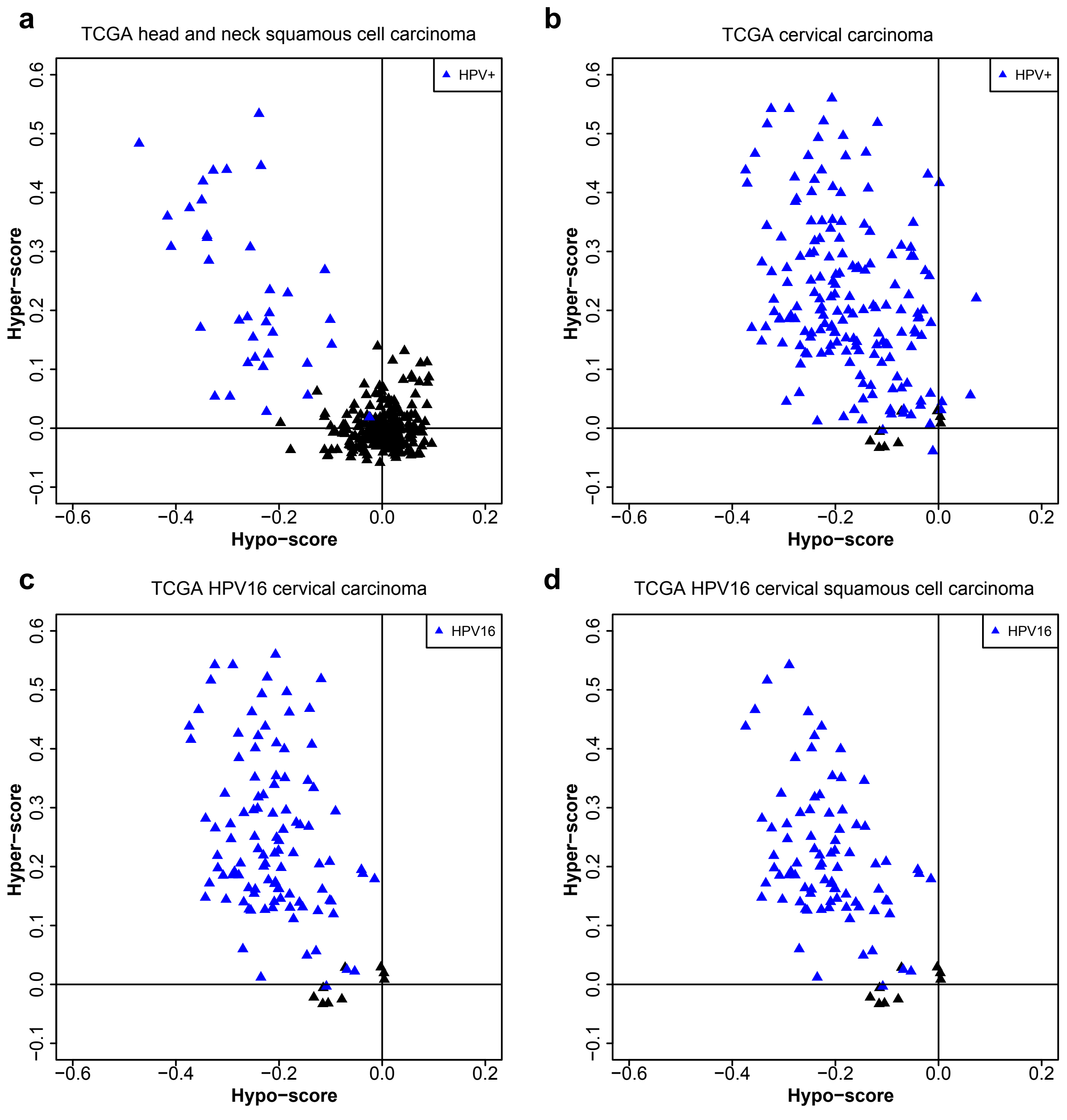
Supplemental Figure S8: Survival analysis between CN clusters. Kaplan-Meier survival analysis between cases within the CN High (High) and CN Low (Low) clusters.



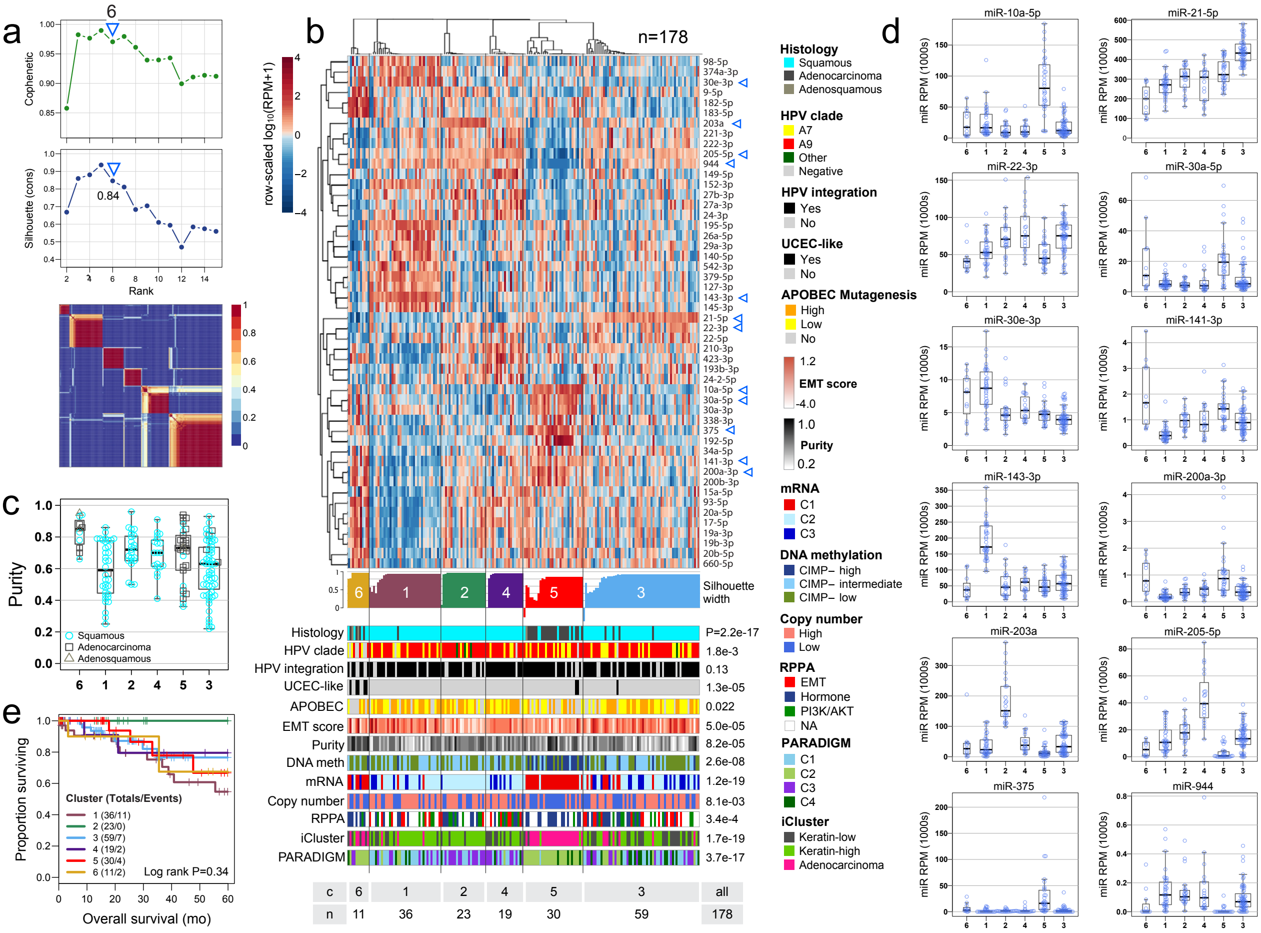
Supplemental Figure S9: **mRNA clustering analysis**. Gene expression values obtained from RNAseq data on 300 signature genes (y-axis) across 178 cervical cancer samples (x-axis) were hierarchically clustered using uncentered correlation and centroid linkage as the clustering method (left). Normalized gene expression values were median-centered prior to clustering and relative increased expression values are indicated by red color while relative decreased expression values are indicated by blue color. Sample annotations are indicated above the sample dendrogram. Select genes are noted to the right of their locations on the heatmap. Horizonal black lines approximately separate gene clusters. Five-year survival analysis of cervical cancer patients grouped according to mRNA cluster membership (lower right panel).


3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
COL7A1
FAT2
S100A8, S100A9
KRT17
IL8
MAPK10
THBS4
PCDHB2
TRO
CADM1
CAPS
ESR1
PGR
Tumor Type
Endometrial
Cervical
*
**
*
**
**
*: UCEC-like samples

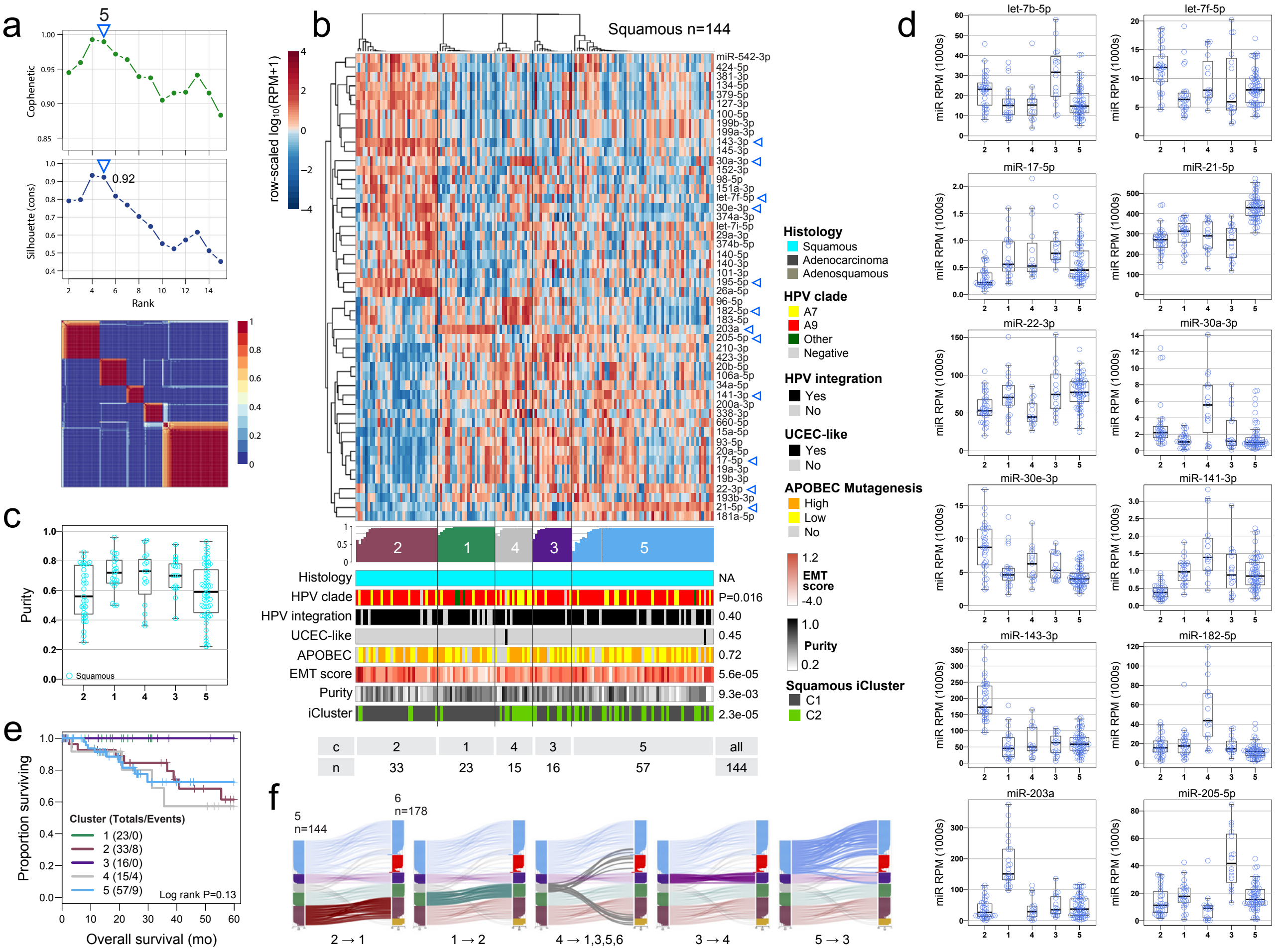
Supplemental Figure S10: **Clustering analysis of CESC and UCEC TCGA samples**. ANOVA was performed on normalized gene-level RSEM values for 178 cervical cancer and 170 TCGA endometrial cancer samples to identify differentially expressed genes between the two cancer types. The differentially expressed genes (n=384, FDR <0.05) and samples were clustered using uncentered correlation and centroid linkage as the clustering method. RSEM normalized values were median-centered prior to clustering and relative increased expression values are indicated by red color while relative decreased expression values are indicated by blue color. Cervical and endometrial cancer samples are indicated by different colors as noted in the figure, and the 8 endometrial-like (UCEC-like) cervical cancer samples are noted with an *. Select genes are noted to the right of their locations on the heatmap.



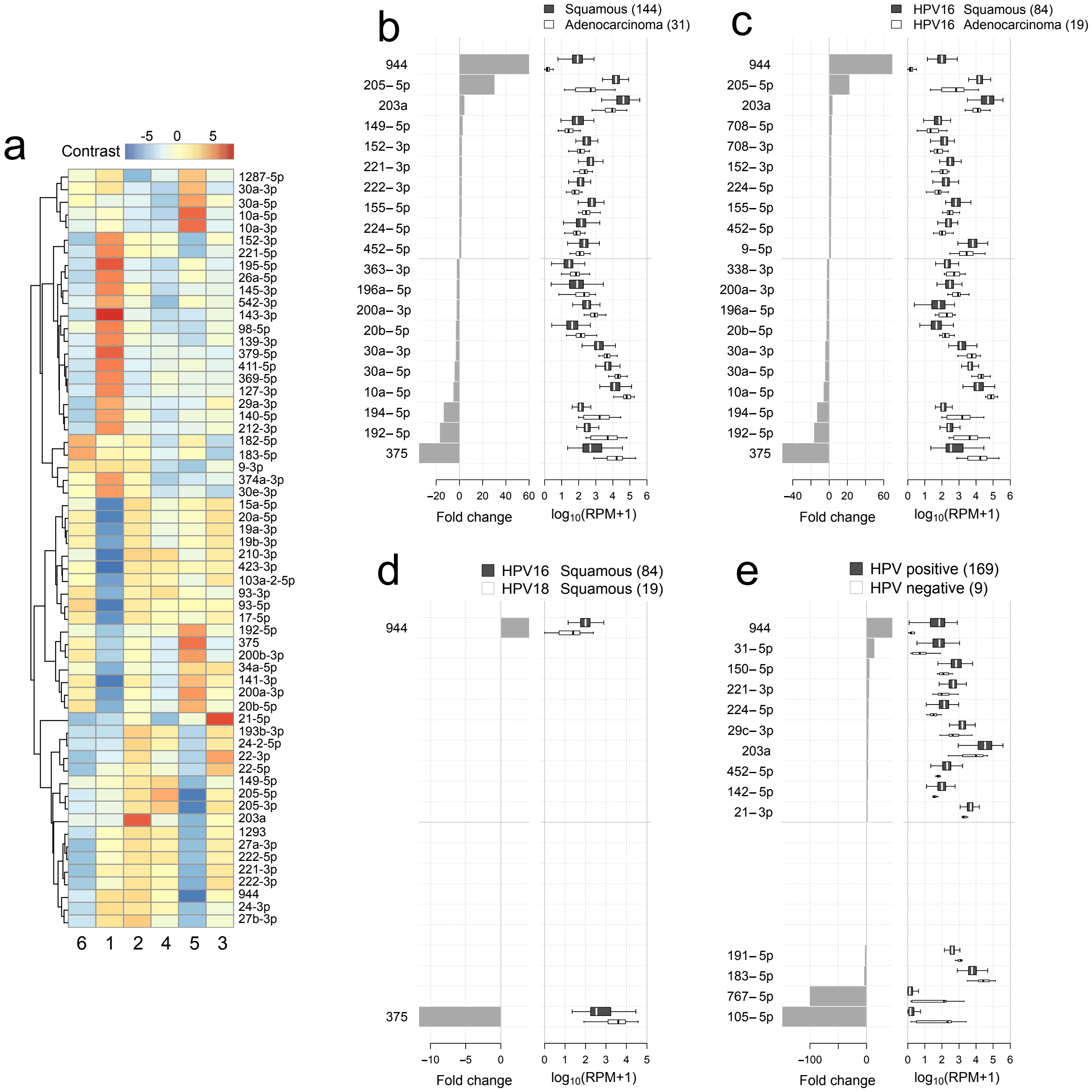
Supplemental Figure S11: DNA methylation signatures of HPV derived on TCGA head and neck cancer sample cohort. **a-d**, The distribution of DNA hyper- and hypo-methylation scores for a) TCGA head and neck squamous cell carcinoma samples, b) all cervical carcinoma samples, c) HPV16 squamous cell carcinomas and adenocarcinomas of the cervix, and d) HPV16 cervical squamous cell carcinomas. HPV positive samples are shown in blue and HPV negative samples are shown in black.


a
6
Cophenetic
Silhouette (cons)
0.84
Rank
b
n=178
row-scaled log10(RPM+1)
98-5p
374a-3p
30e-3p
9-5p
182-5p
183-5p
203a
221-3p
222-3p
205-5p
944
149-5p
152-3p
27b-3p
27a-3p
24-3p
195-5p
26a-5p
29a-3p
140-5p
542-3p
379-5p
127-3p
143-3p
145-3p
21-5p
22-3p
22-5p
210-3p
423-3p
193b-3p
24-2-5p
10a-5p
30a-5p
30a-3p
338-3p
375
192-5p
34a-5p
141-3p
200a-3p
200b-3p
15a-5p
93-5p
20a-5p
17-5p
19a-3p
19b-3p
20b-5p
660-5p
Silhouette width
Histology P=2.2e-17
HPV clade 1.8e-3
HPV integration 0.13
UCEC-like 1.3e-05
APOBEC 0.022
EMT score 5.0e-05
Purity 8.2e-05
DNA meth 2.6e-08
mRNA 1.2e-19
Copy number 8.1e-03
RPPA 3.4e-4
iCluster 1.7e-19
PARADIGM 3.7e-17
c 6 1 2 4 5 3 all
n 11 36 23 19 30 59 178
Histology: Squamous, Adenocarcinoma, Adenosquamous
HPV clade: A7, A9, Other, Negative
HPV integration: Yes, No
UCEC-like: Yes, No
APOBEC Mutagenesis: High, Low, No
EMT score 1.2 -4.0
Purity 1.0 0.2
mRNA: C1, C2, C3
DNA methylation: CIMP– high, CIMP– intermediate, CIMP– low
Copy number: High, Low
RPPA: EMT, Hormone, PI3K/AKT, NA
PARADIGM: C1, C2, C3, C4
iCluster: Keratin-low, Keratin-high, Adenocarcinoma
c
Purity
Squamous
Adenocarcinoma
Adenosquamous
d
miR-10a-5p
miR-21-5p
miR-22-3p
miR-30a-5p
miR-30e-3p
miR-141-3p
miR-143-3p
miR-200a-3p
miR-203a
miR-205-5p
miR-375
miR-944
miR RPM (1000s)
e
Proportion surviving
Cluster (Totals/Events)
1 (36/11)
2 (23/0)
3 (59/7)
4 (19/2)
5 (30/4)
6 (11/2)
Log rank P=0.34
Overall survival (mo)

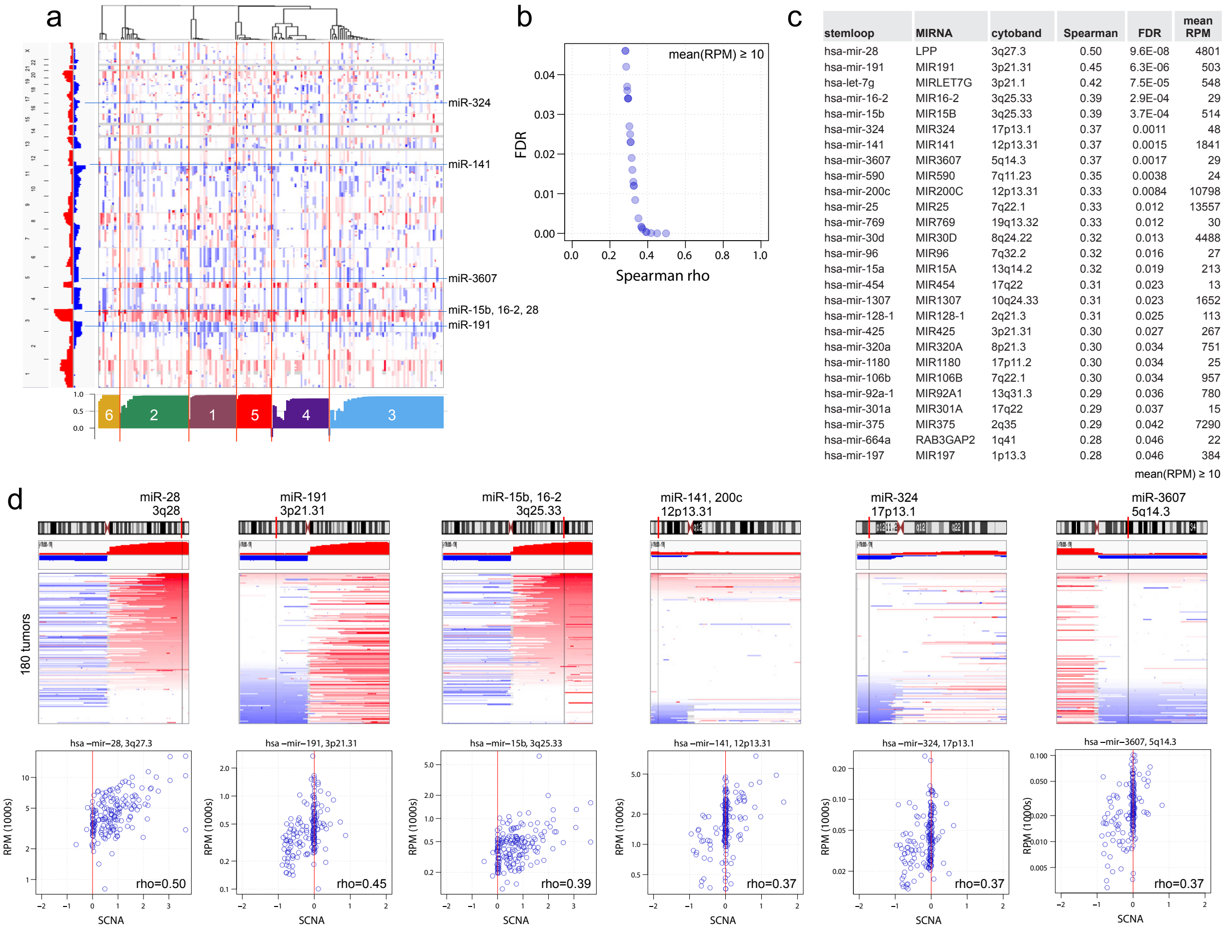
Supplemental Figure S12: Unsupervised NMF consensus clustering of miRNA mature strand data. **a,** Rank survey profiles for cophenetic correlation coefficient and average silhouette width for the NMF consensus clustering rank survey for 178 tumor samples, and a blue/red heatmap showing sample consensus memberships for a six-cluster solution, with yellow-white indicating samples that are less 'typical' cluster members. **b**, For the six-cluster solution, top to bottom: a normalized abundance heatmap for the fifty 5p or 3p strands that were highly ranked as differentially abundant by a SAMseq multiclass analysis, silhouette width profile calculated from the consensus membership matrix, covariates with Fisher exact association p-values, and a summary table of cluster number and the number of samples in each cluster. The scale bar shows row-scaled log10(RPM+1) normalized abundances. **c**, Per-cluster distributions of tumor purity. **d**, Per-cluster distributions of normalized (RPM) abundance for a subset of miRs that were differentially abundant across the unsupervised clusters and had relatively high RPMs. Black horizontal bars indicate median RPMs. **e**, Kaplan-Meier plot of overall survival.



Supplemental Figure S13: Unsupervised NMF consensus clustering of miRNA mature strand data for squamous tumors (n=144). a-e, Panels are as in Supplemental Figure S12. **f,** Relationships between sample locations across the current 5-cluster solution for n=144, and the 6-cluster solution for n=178 in Supplemental Fig. S12. In each of the five graphics, each curve shows the location of a sample in the two clustering solutions, and curves for all samples in one of the n=144 clusters are highlighted. Text below a graphic summarizes the clusters that the squamous samples segregate to in the all sample clustering solution from each cluster of the squamous sample solution. Clusters that the curves indicate have similar sample memberships in the n=178 and n=144 solutions are assigned the same color in **b** and **f**.


a
Contrast
-5
0
5
1287-5p
30a-3p
30a-5p
10a-5p
10a-3p
152-3p
221-5p
195-5p
26a-5p
145-3p
542-3p
143-3p
98-5p
139-3p
379-5p
411-5p
369-5p
127-3p
29a-3p
140-5p
212-3p
182-5p
183-5p
9-3p
374a-3p
30e-3p
15a-5p
20a-5p
19a-3p
19b-3p
210-3p
423-3p
103a-2-5p
93-3p
93-5p
17-5p
192-5p
375
200b-3p
34a-5p
141-3p
200a-3p
20b-5p
21-5p
193b-3p
24-2-5p
22-3p
22-5p
149-5p
205-5p
205-3p
203a
1293
27a-3p
222-5p
221-3p
222-3p
944
24-3p
27b-3p
6 1 2 4 5 3
b
Squamous (144)
Adenocarcinoma (31)
944
205–5p
203a
149–5p
152–3p
221–3p
222–3p
155–5p
224–5p
452–5p
363–3p
196a–5p
200a–3p
20b–5p
30a–3p
30a–5p
10a–5p
194–5p
192–5p
375
–20 0 20 40 60
Fold change
0 1 2 3 4 5 6
log10(RPM+1)
c
HPV16 Squamous (84)
HPV16 Adenocarcinoma (19)
944
205–5p
203a
708–5p
708–3p
152–3p
224–5p
155–5p
452–5p
9–5p
338–3p
200a–3p
196a–5p
20b–5p
30a–3p
30a–5p
10a–5p
194–5p
192–5p
375
–40 0 20 60
Fold change
0 1 2 3 4 5 6
log10(RPM+1)
d
HPV16 Squamous (84)
HPV18 Squamous (19)
944
375
–10 –5 0
Fold change
0 1 2 3 4 5
log10(RPM+1)
e
HPV positive (169)
HPV negative (9)
944
31–5p
150–5p
221–3p
224–5p
29c–5p
203a
452–5p
142–5p
21–3p
191–5p
183–5p
767–5p
105–5p
–100 0
Fold change
0 1 2 3 4 5 6
log10(RPM+1)

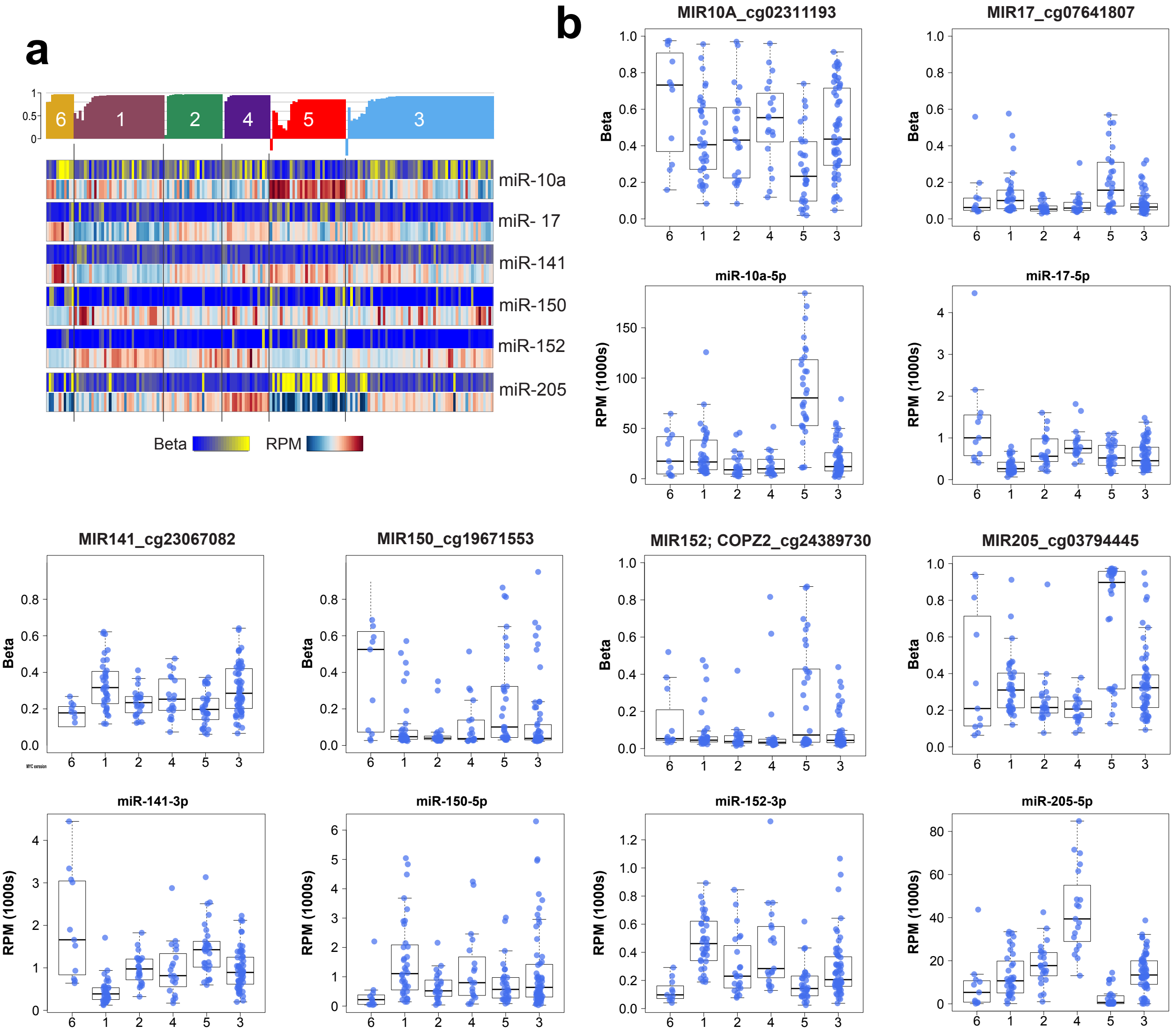
Supplemental Figure S14: Differentially abundant miRs. **a,** Heatmap of differential abundance contrasts for the 50 miRs that were scored highly by a SAMseq multiclass analysis across the six unsupervised clusters. **b-e,** miRs with the largest fold-changes for b) squamous vs. adenocarcinomas, c) HPV16 squamous vs. HPV16 adenocarcinomas, d) HPV16 squamous vs. HPV18 squamous carcinomas, and e) HPV positive (+) vs. HPV negative (-) tumors. Each panel has (left) a barplot of median-based fold-change, and (right) boxplots showing distributions of normalized (RPM) abundance, with black/white vertical lines indicating medians. Up to 10 of the largest fold-changes in each direction are shown. The numbers of samples in each group are in parentheses. miRs that have a mean abundance of at least 50 RPM are presented in each graph.



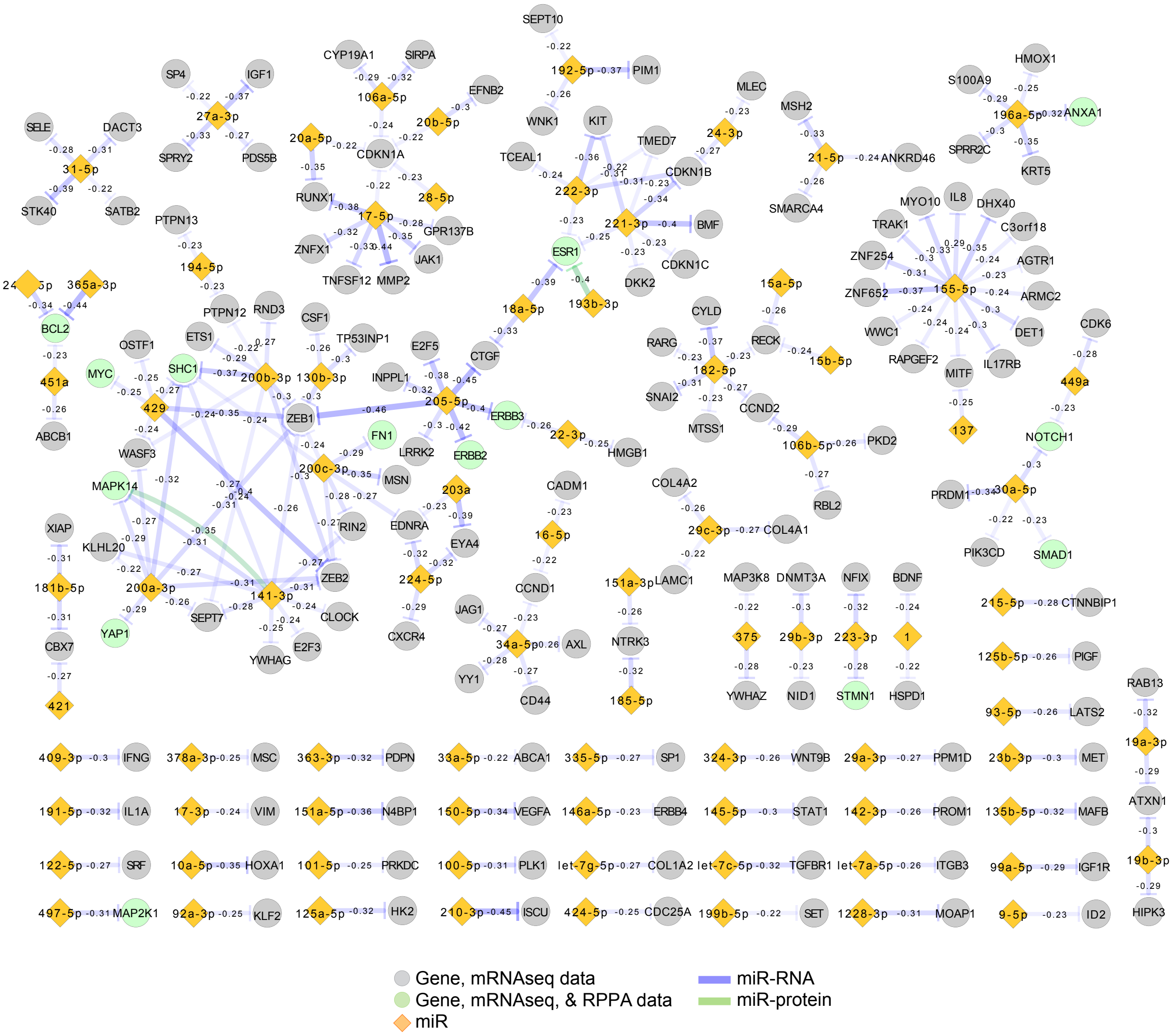
stemloop	MIRNA	cytoband	Spearman	FDR	mean RPM
hsa-mir-28	LPP	3q27.3	0.50	9.6E-08	4801
hsa-mir-191	MIR191	3p21.31	0.45	6.3E-06	503
hsa-let-7g	MIRLET7G	3p21.1	0.42	7.5E-05	548
hsa-mir-16-2	MIR16-2	3q25.33	0.39	2.9E-04	29
hsa-mir-15b	MIR15B	3q25.33	0.39	3.7E-04	514
hsa-mir-324	MIR324	17p13.1	0.37	0.0011	48
hsa-mir-141	MIR141	12p13.31	0.37	0.0015	1841
hsa-mir-3607	MIR3607	5q14.3	0.37	0.0017	29
hsa-mir-590	MIR590	7q11.23	0.35	0.0038	24
hsa-mir-200c	MIR200C	12p13.31	0.33	0.0084	10798
hsa-mir-25	MIR25	7q22.1	0.33	0.012	13557
hsa-mir-769	MIR769	19q13.32	0.33	0.012	30
hsa-mir-30d	MIR30D	8q24.22	0.32	0.013	4488
hsa-mir-96	MIR96	7q32.2	0.32	0.016	27
hsa-mir-15a	MIR15A	13q14.2	0.32	0.019	213
hsa-mir-454	MIR454	17q22	0.31	0.023	13
hsa-mir-1307	MIR1307	10q24.33	0.31	0.023	1652
hsa-mir-128-1	MIR128-1	2q21.3	0.31	0.025	113
hsa-mir-425	MIR425	3p21.31	0.30	0.027	267
hsa-mir-320a	MIR320A	8p21.3	0.30	0.034	751
hsa-mir-1180	MIR1180	17p11.2	0.30	0.034	25
hsa-mir-106b	MIR106B	7q22.1	0.30	0.034	957
hsa-mir-92a-1	MIR92A1	13q31.3	0.29	0.036	780
hsa-mir-301a	MIR301A	17q22	0.29	0.037	15
hsa-mir-375	MIR375	2q35	0.29	0.042	7290
hsa-mir-664a	RAB3GAP2	1q41	0.28	0.046	22
hsa-mir-197	MIR197	1p13.3	0.28	0.046	384

mean(RPM) ≥ 10

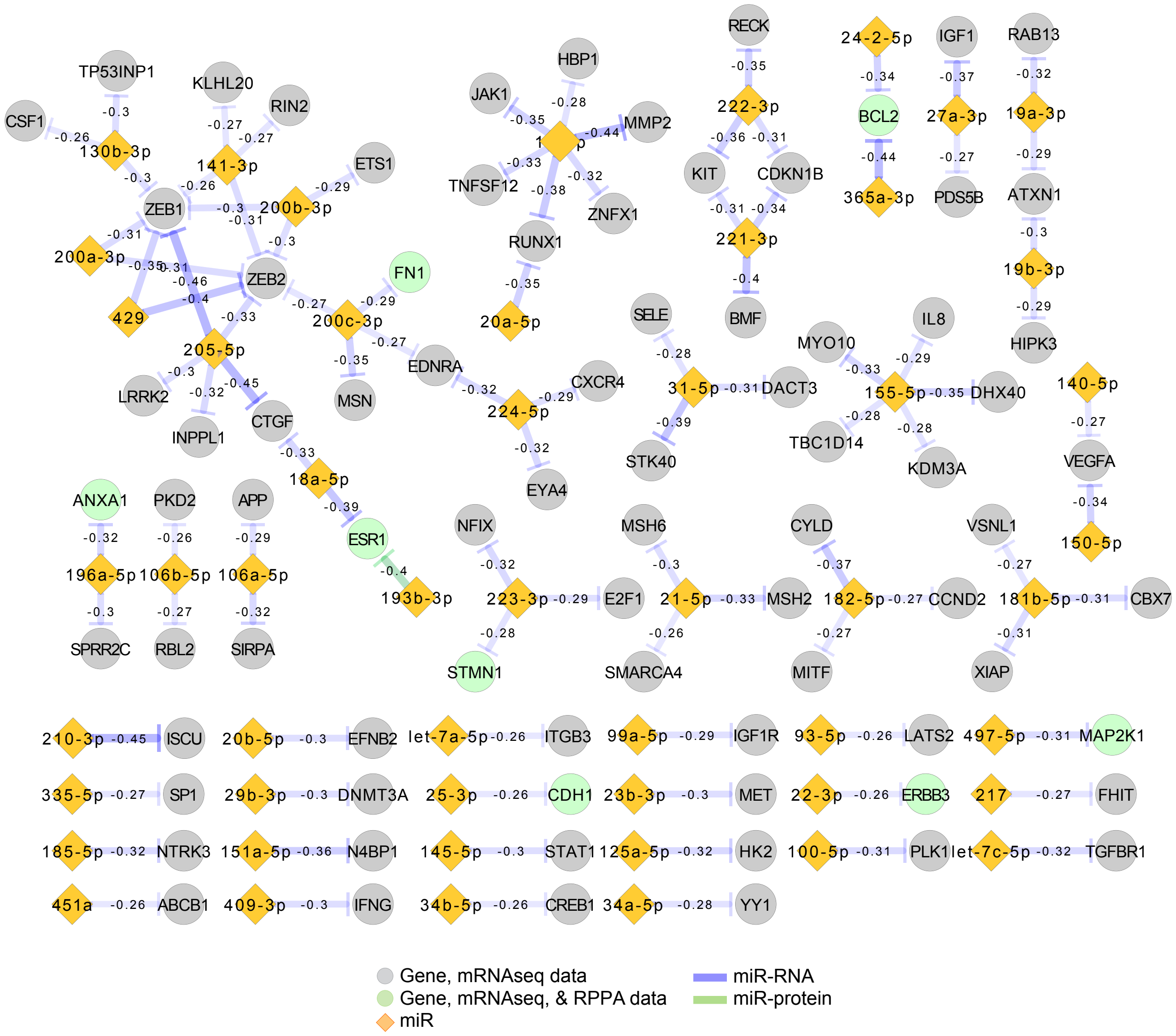
Supplemental Figure S15: Relationships between somatic copy number (SCNA) and pre-miRNA abundance. **a,** Global relationship between SCNA and miR-based unsupervised clusters. Blue horizontal lines mark the locations of the miRNAs in panel **d**. **b**, Relationship between the Spearman correlation coefficient (rho) and correlation false discovery rate (FDR). **c,** All pre-miRNAs (stemloops) with correlation FDR < 0.05. **d,** Details for six example stemloops whose correlations are statistically significant in **c**. For each miRNA, the upper graphic shows SCNA for a chromosome sorted by amplification at the miRNA's location, and the scatterplot shows the relationship between SCNA and pre-miRNA normalized abundance (RPM), with the Spearman correlation coefficient presented in the lower right corner.



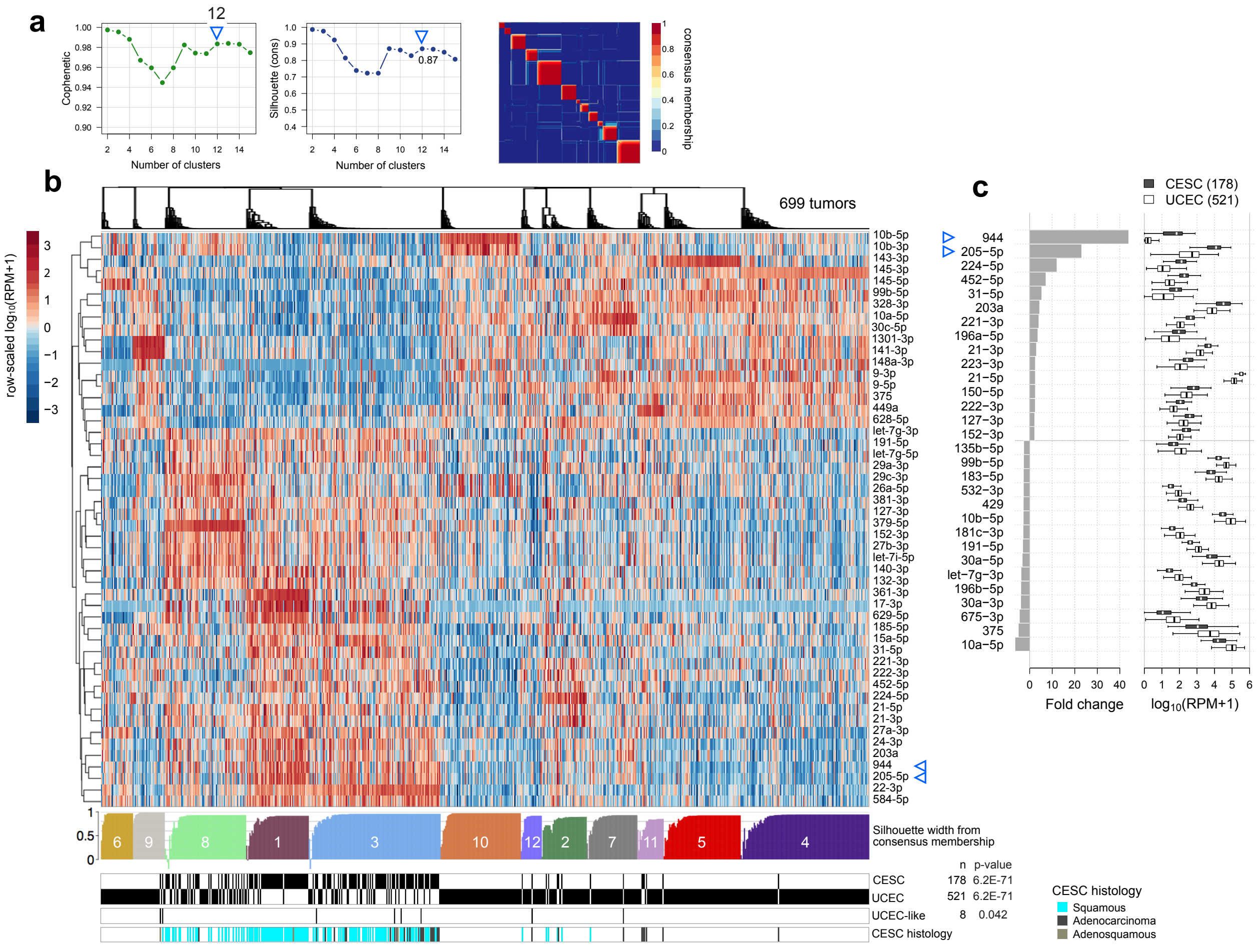
Supplemental Figure S16: miRNAs that may be influenced by DNA methylation. a, Covariate tracks showing beta values for DNA methylation and stemloop normalized abundance (RPM). **b,** Distributions across miR-based clusters (all samples) of methylation beta for a correlated DNA methylation probe (above) and stemloop abundance (below).



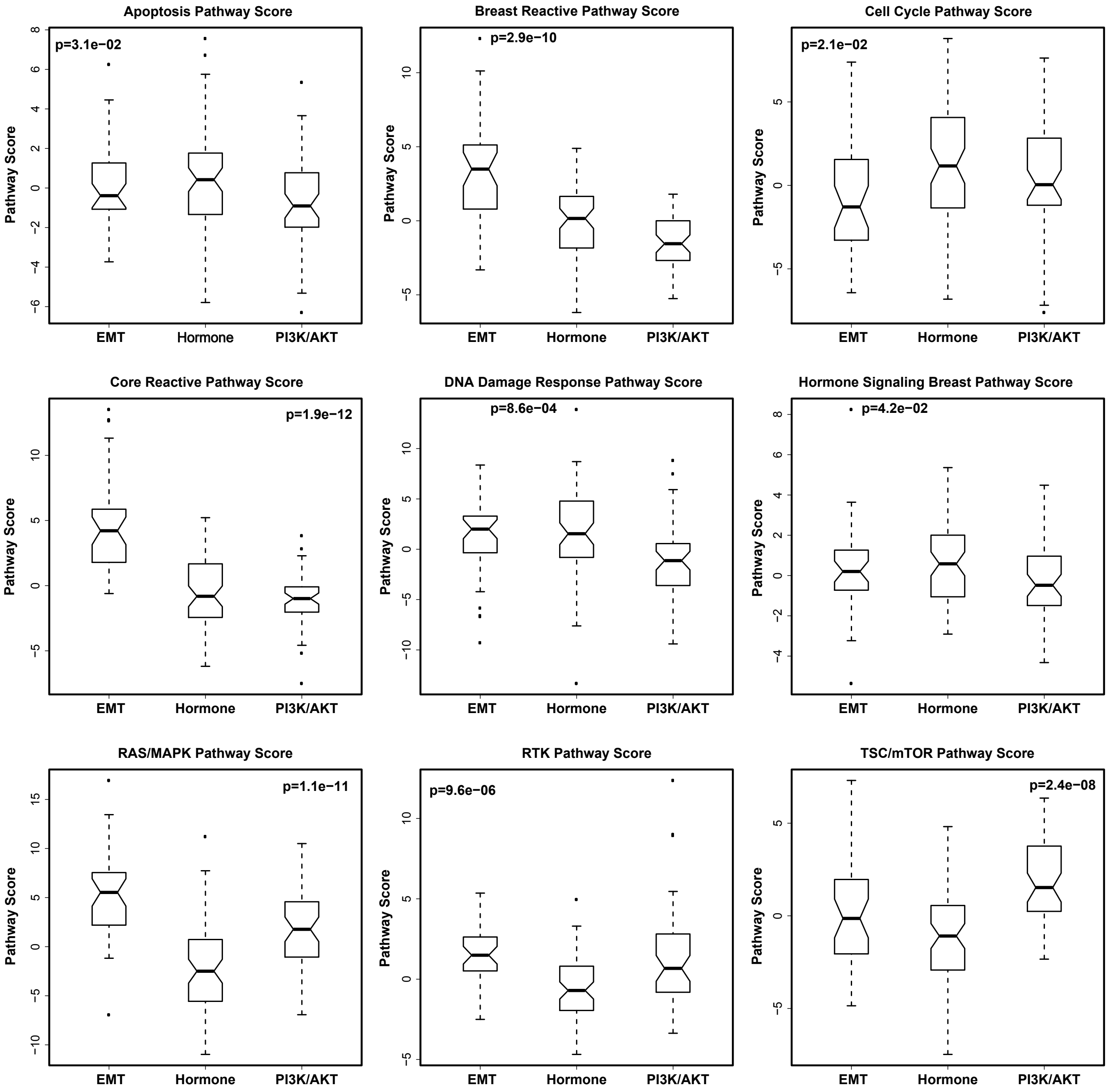
Supplemental Figure S17: Functionally validated potential miR-gene and miR-protein targeting. Significance-thresholded (FDR<0.05) miR-mRNA anti-correlations that are supported by functional validation publications with strong evidence types. For genes, node color distinguishes those that are only present in mRNA data (grey) from those that are present in both mRNA and RPPA data (green). Edges represent anti-correlations, and color distinguishes anti-correlations between a miR and mRNA (purple) and a miR and an unphosphorylated protein (green). In the all samples cohort, no correlations satisfying FDR<0.05 were reported between a miR and a phosphorylated protein.



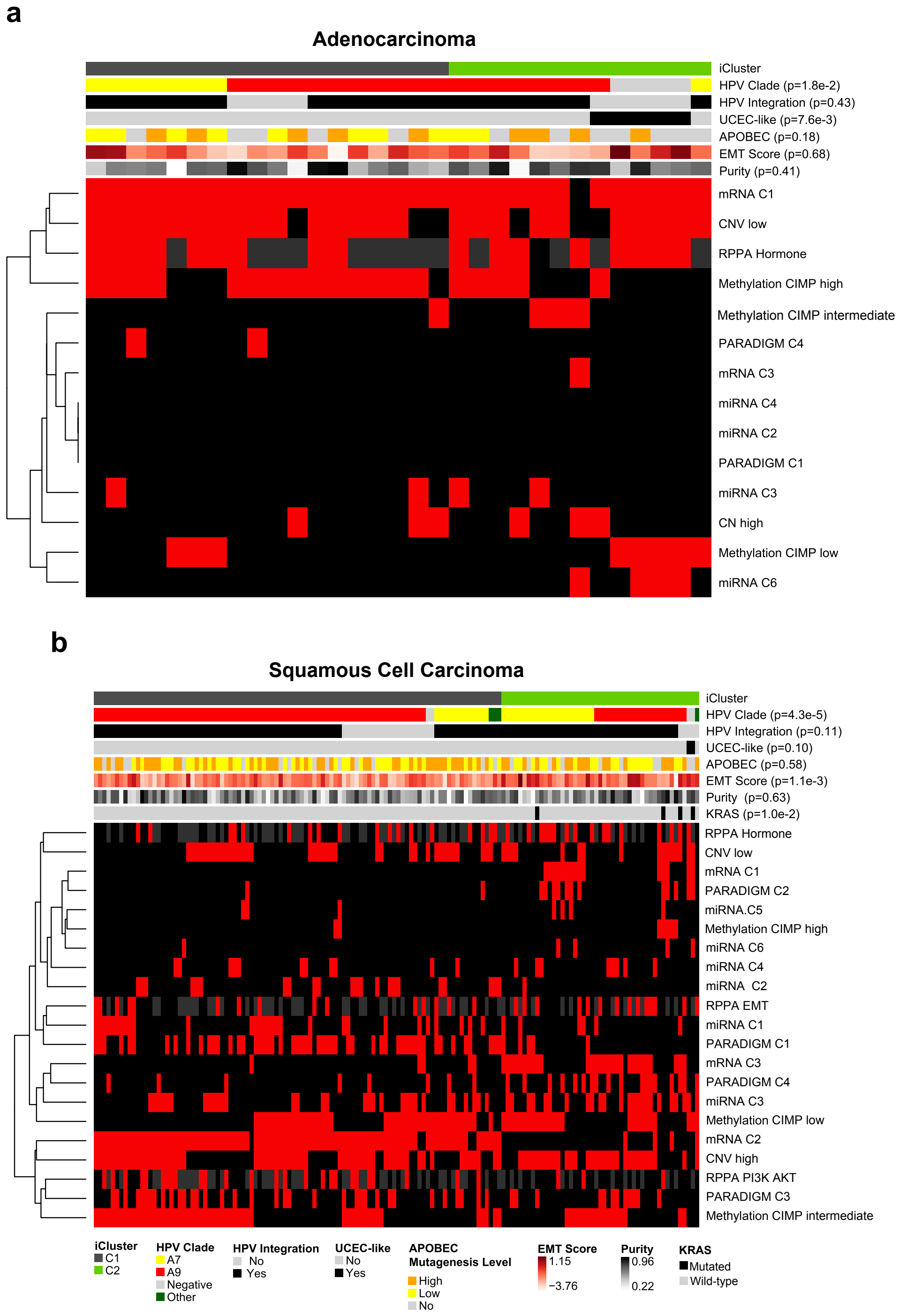
Supplemental Figure S18: Functionally validated potential miR-gene and miR-RPPA targeting for squamous samples. See legend of Supplemental Figure S17.



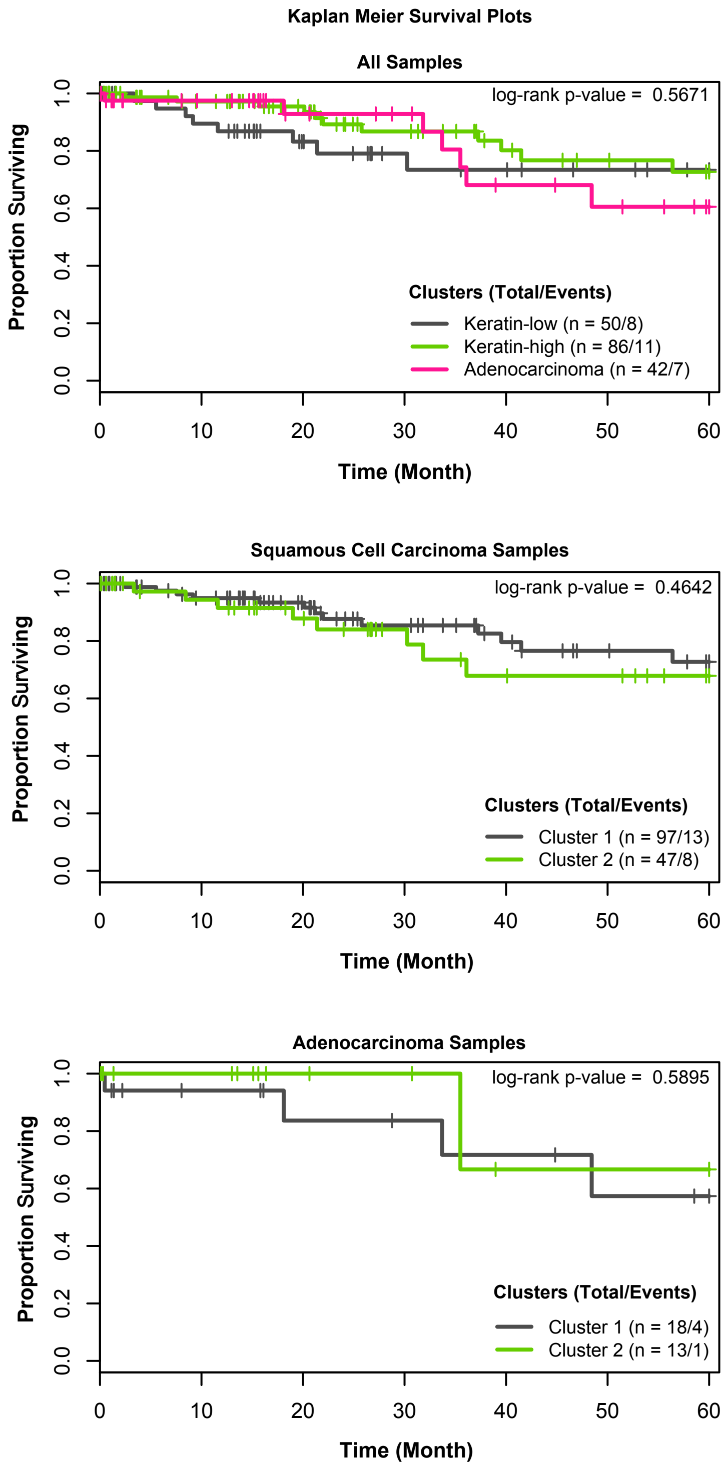
Supplemental Figure S19: Unsupervised clustering and differentially abundant miRs for 521 endometrial and 178 cervical tumor samples. **a,** Profiles of cophenetic correlation coefficient and silhouette width calculated from the consensus membership for solutions with 2 to 15 clusters. The red-blue heatmap shows consensus membership values for the 12-cluster solution. **b,** Top to bottom: A row-scaled, normalized abundance heatmap for the 50 miRs scored most highly in a multiclass SAMSeq analysis, a silhouette width profile, and covariate tracks for disease type, endometrial-like (UCEC-like) CESC samples, and the three CESC histological types. **c,** Differentially abundant miRs between UCEC and CESC samples (FDR<0.05). Triangles in **b** and **c** highlight miR-944 and miR-205.



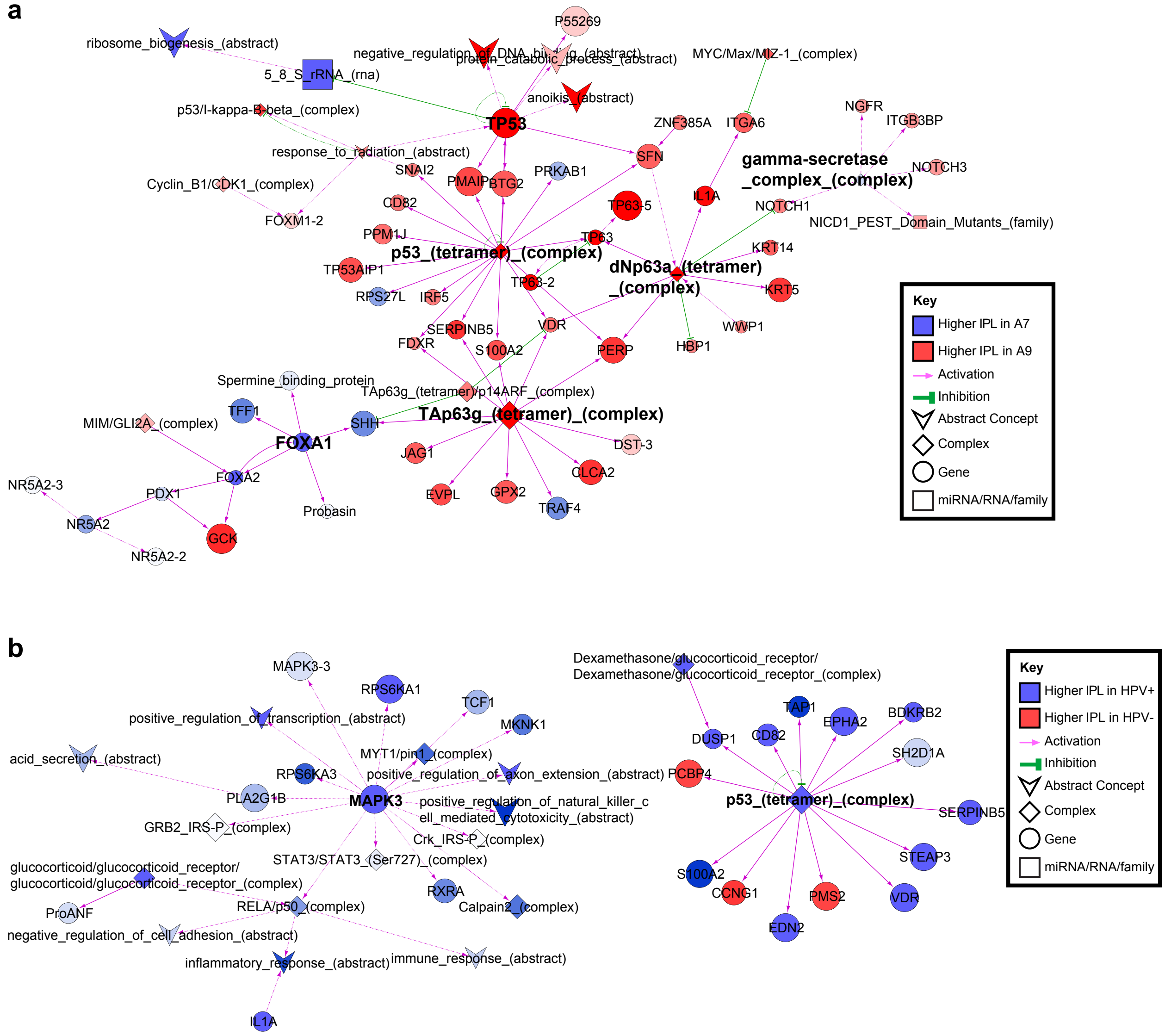
Supplemental Figure S20: Protein signaling pathway scores for all samples. Pathway scores for apoptosis, reactive breast, cell cycle, core reactive, DNA damage response, breast hormone, RAS/MAPK, RTK, and TSC/mTOR signaling pathways are presented with significant pathway score differences between the clusters measured by Kruskal Wallis test.


a
Adenocarcinoma
iCluster
HPV Clade (p=1.8e-2)
HPV Integration (p=0.43)
UCEC-like (p=7.6e-3)
APOBEC (p=0.18)
EMT Score (p=0.68)
Purity (p=0.41)
mRNA C1
CNV low
RPPA Hormone
Methylation CIMP high
Methylation CIMP intermediate
PARADIGM C4
mRNA C3
miRNA C4
miRNA C2
PARADIGM C1
miRNA C3
CN high
Methylation CIMP low
miRNA C6
b
Squamous Cell Carcinoma
iCluster
HPV Clade (p=4.3e-5)
HPV Integration (p=0.11)
UCEC-like (p=0.10)
APOBEC (p=0.58)
EMT Score (p=1.1e-3)
Purity (p=0.63)
KRAS (p=1.0e-2)
RPPA Hormone
CNV low
mRNA C1
PARADIGM C2
miRNA.C5
Methylation CIMP high
miRNA C6
miRNA C4
miRNA C2
RPPA EMT
miRNA C1
PARADIGM C1
mRNA C3
PARADIGM C4
miRNA C3
Methylation CIMP low
mRNA C2
CNV high
RPPA PI3K AKT
PARADIGM C3
Methylation CIMP intermediate
iCluster
C1
C2
HPV Clade
A7
A9
Negative
Other
HPV Integration
No
Yes
UCEC-like
No
Yes
APOBEC Mutagenesis Level
High
Low
No
EMT Score
1.15
−3.76
Purity
0.96
0.22
KRAS
Mutated
Wild-type

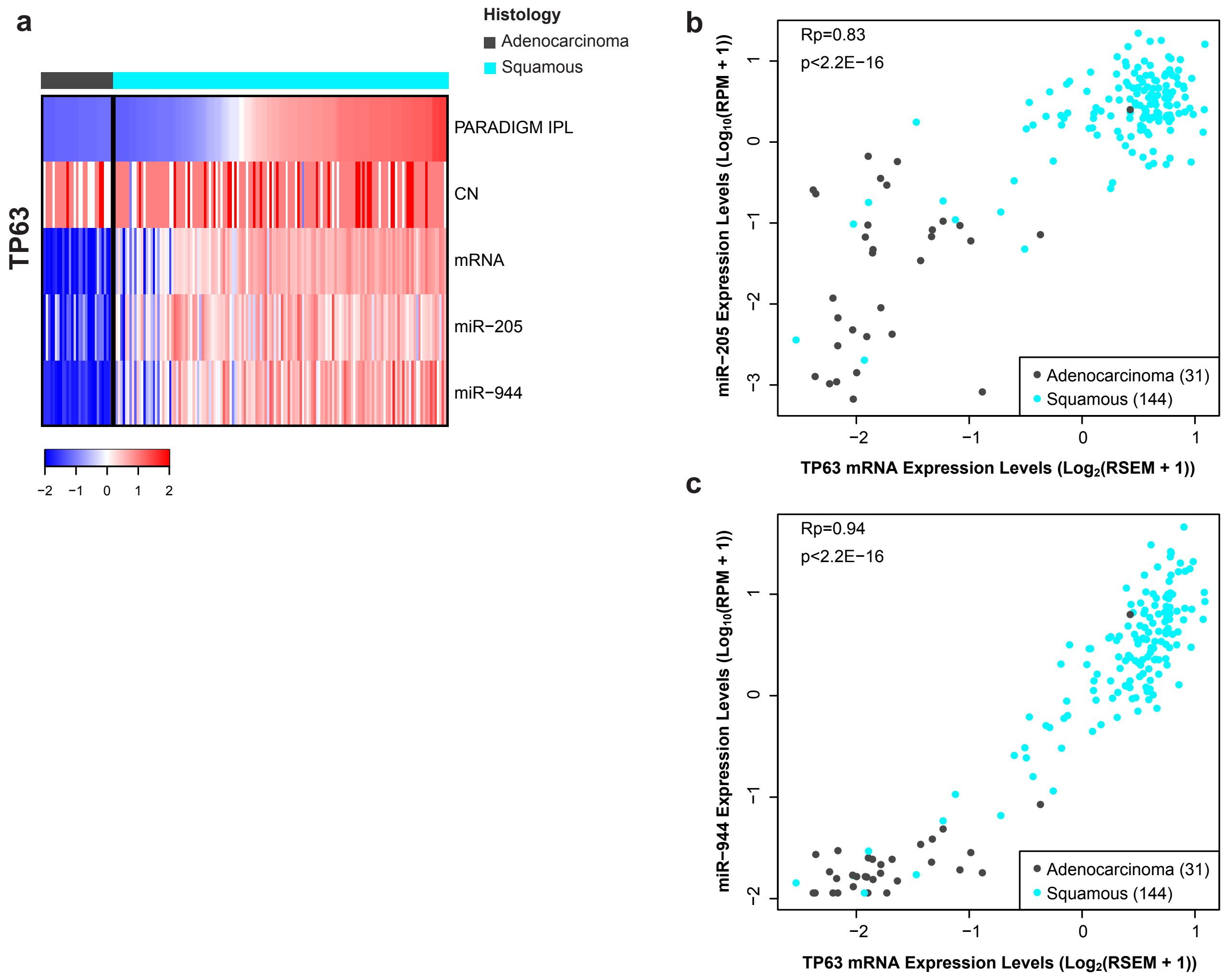
Supplemental Figure S21: Integrative clustering of cervical squamous cell and adenocarcinomas. a-b, Integrative clustering of 31 cervical adenocarcinomas (a) and 144 squamous cell carcinomas (b) using mRNA, methylation, miRNA, and CNV data. The feature bars at the top show the iCluster, HPV clade, HPV integration status, UCEC-like status, APOBEC mutagenesis level, mRNA EMT score, and tumor purity. The mutation status of the SMG *KRAS* is also shown in the squamous iClusters. The cluster of cluster panel displays subtypes defined independently by mRNA, miRNA, methylation, reverse phase protein array (RPPA), copy number (CNV), and PARADIGM. Platform clusters that did not contain adenocarcinoma samples are excluded from the adenocarcinoma panel. Black indicates that the sample is not represented in the cluster, red indicates that the sample is represented in the cluster, and gray represents data not available.



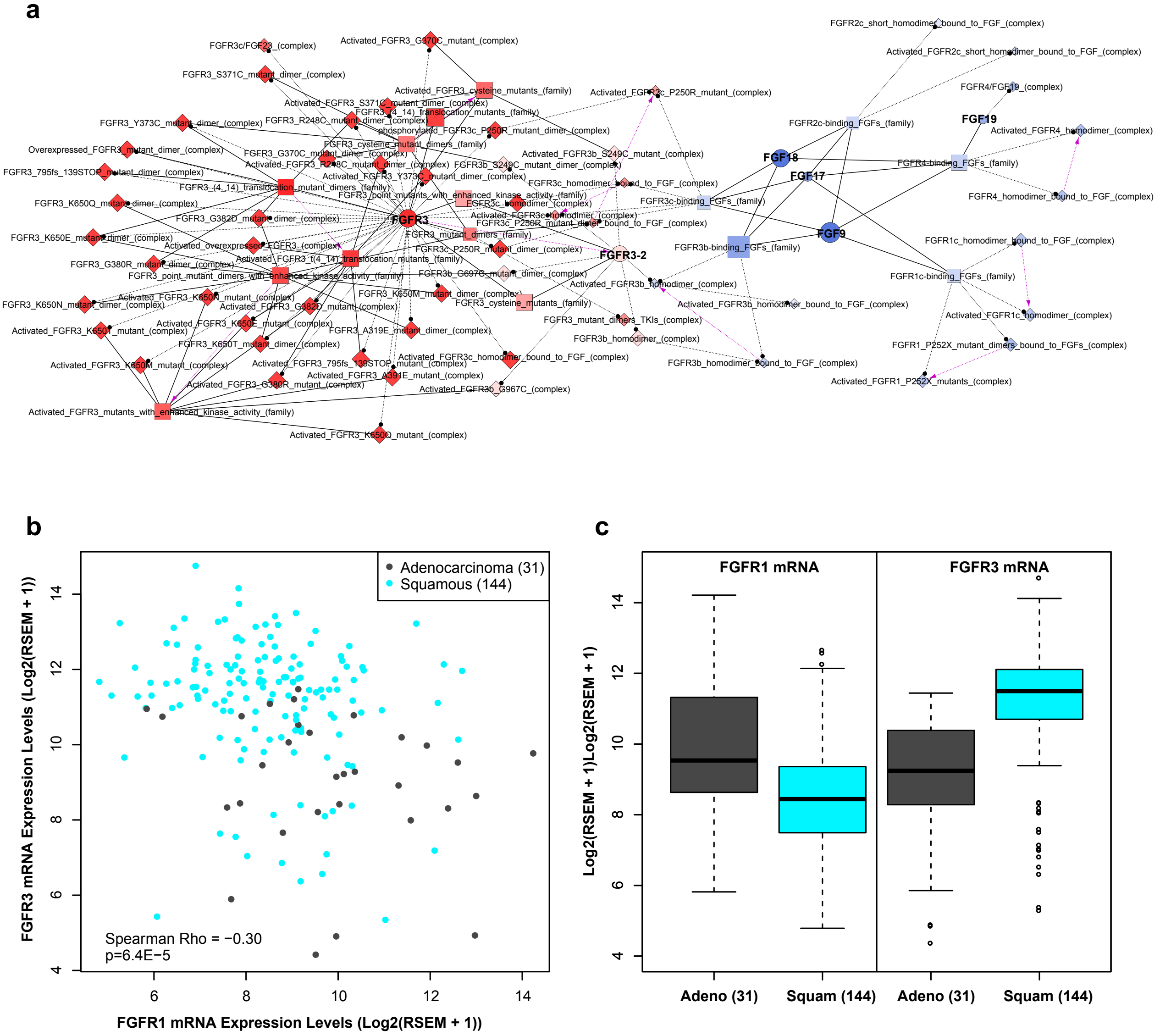
Supplemental Figure S22: Survival analysis across iClusters. Kaplan Meier survival curves assessing survival differences among the three clusters of all histology combined (top) and between the two clusters of both squamous (middle) and adenocarcinoma samples (bottom) with p-values calculated from log- rank test.



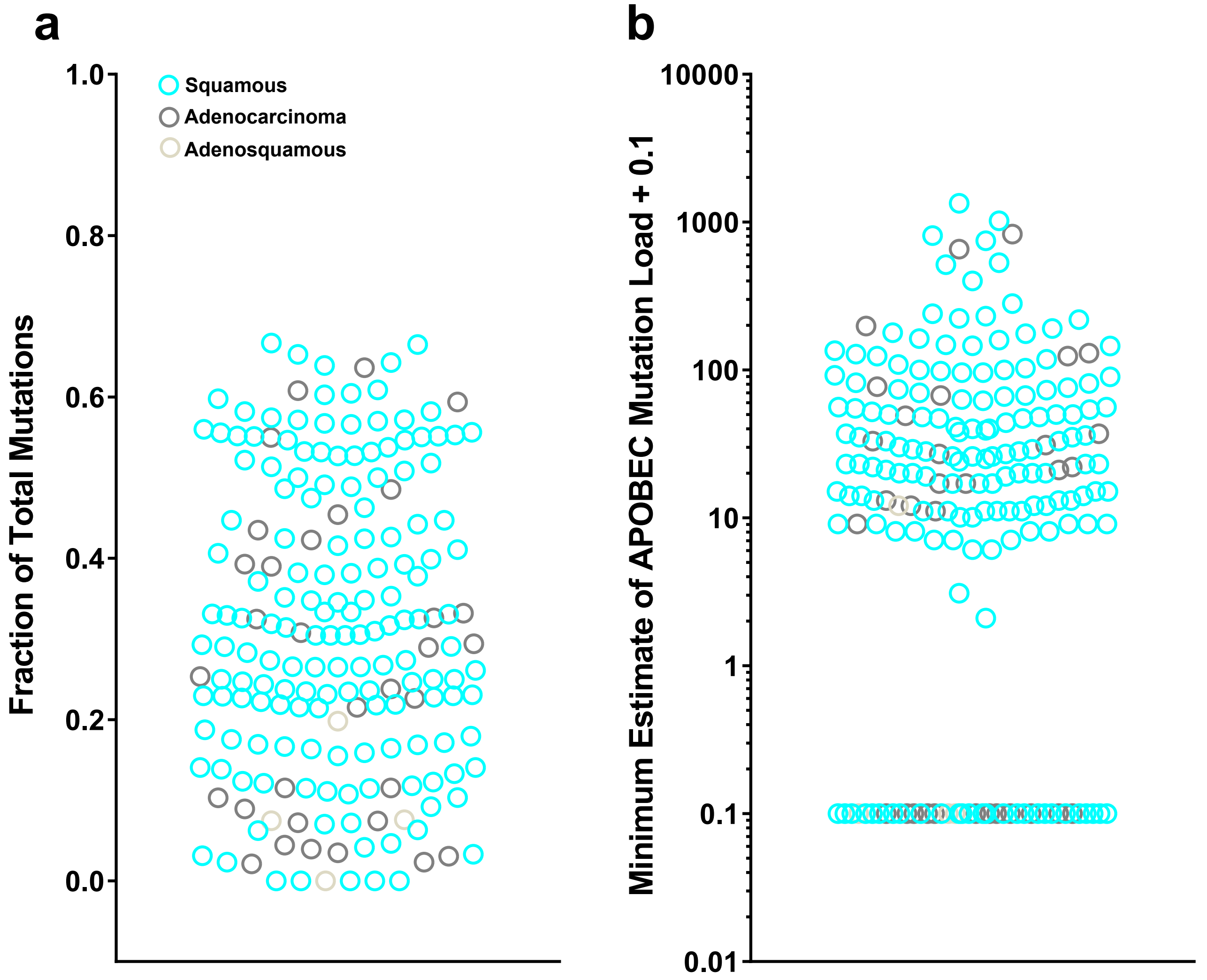
Supplemental Figure S23: Pathway biomarkers associated with HPV status. **a,** Cytoscape display of the largest interconnected regulatory network of features differentially activated between HPV A9 and A7 positive cervical cancers of all histologies. **b,** Cytoscape display of the two largest interconnected regulatory networks of features differentially activated between HPV negative and HPV positive cervical cancers.



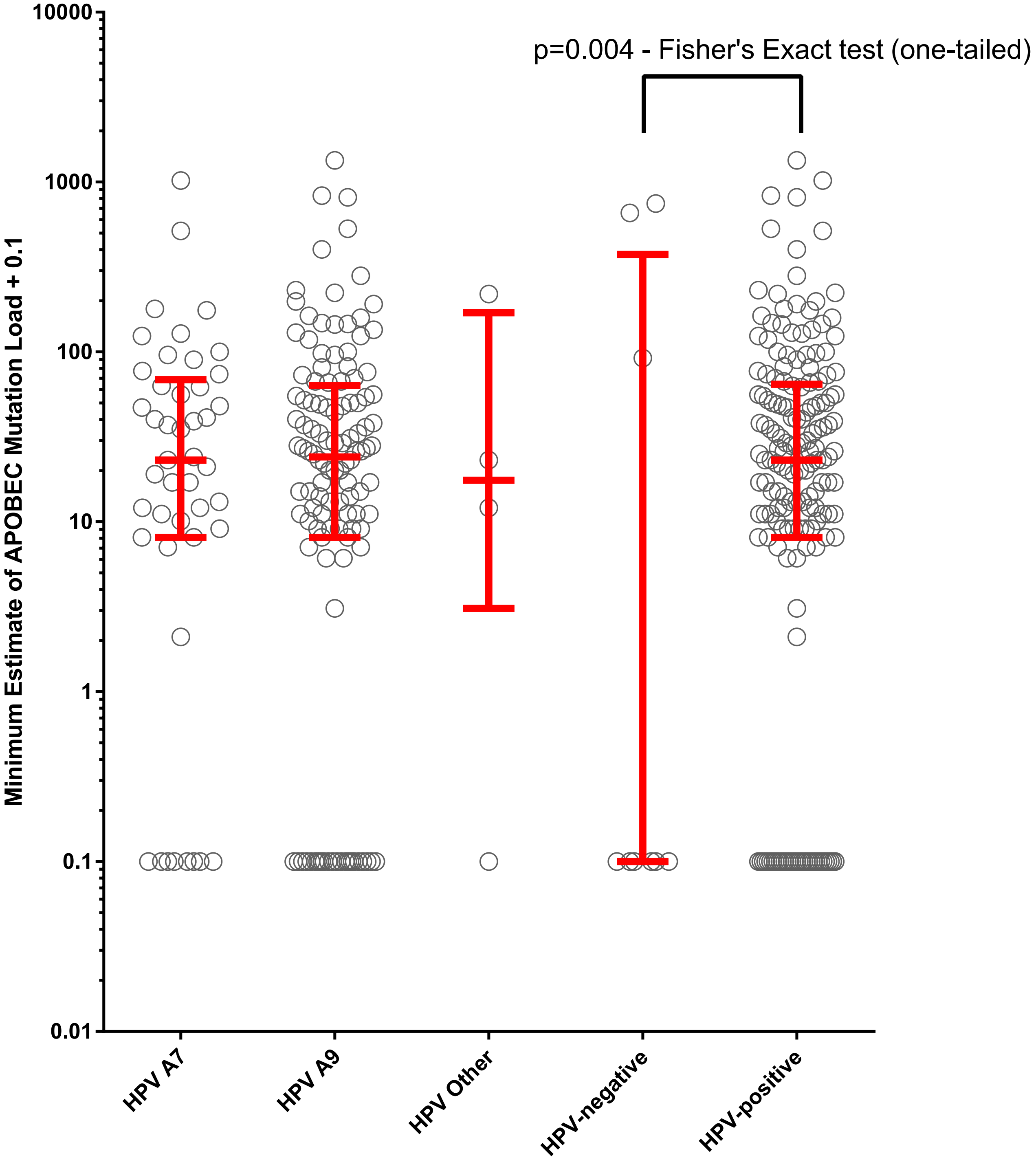
Supplemental Figure S24: p63 as a marker distinguishing squamous carcinomas and adenocarcinomas. **a,** Heatmap showing p63 PARADIGM IPL, copy number (CN), and mRNA expression levels, as well as miR-944 (which is located in the intron of p63) and miR-205 (which has previously been shown to be p63-regulated) expression levels. Scale represents median-centered IPL, scaled to mean 0 and standard deviation of 1. **b-c,** Scatterplots of p63 mRNA expression vs. miRNA expression levels of miR-205 (b) and miR-944 (c).


a
FGFR3c/FGF23_(complex)
Activated_FGFR3_G370C_mutant_(complex)
FGFR3_S371C_mutant_dimer_(complex)
Activated_FGFR3_cysteine_mutants_(family)
Activated_FGFR3_S371C_mutant_dimer_(complex)
FGFR3_(4_14)_translocation_mutants_(family)
FGFR3_Y373C_mutant_dimer_(complex)
FGFR3_R248C_mutant_dimer_(complex)
phosphorylated_FGFR3c_P250R_mutant_dimer_(complex)
FGFR3_cysteine_mutant_dimers_(family)
Overexpressed_FGFR3_mutant_dimer_(complex)
FGFR3_G370C_mutant_dimer_(complex)
Activated_FGFR3_R248C_mutant_dimer_(complex)
FGFR3b_S249C_mutant_dimer_(complex)
FGFR3_795fs_139STOP_mutant_dimer_(complex)
Activated_FGFR3_Y373C_mutant_dimer_(complex)
FGFR3_(4_14)_translocation_mutant_dimers_(family)
FGFR3_point_mutants_with_enhanced_kinase_activity_(family)
FGFR3_K650Q_mutant_dimer_(complex)
FGFR3c_homodimer_(complex)
FGFR3_G382D_mutant_dimer_(complex)
FGFR3
Activated_FGFR3c_homodimer_(complex)
FGFR3c_P250R_mutant_dimer_bound_to_FGF_(complex)
FGFR3_K650E_mutant_dimer_(complex)
FGFR3_mutant_dimers_(family)
Activated_overexpressed_FGFR3_(complex)
FGFR3c_P250R_mutant_dimer_(complex)
Activated_FGFR3_t(4_14)_translocation_mutants_(family)
FGFR3_G380R_mutant_dimer_(complex)
FGFR3_point_mutant_dimers_with_enhanced_kinase_activity_(family)
FGFR3b_G697C_mutant_dimer_(complex)
Activated_FGFR3_K650N_mutant_(complex)
FGFR3_K650M_mutant_dimer_(complex)
FGFR3_K650N_mutant_dimer_(complex)
Activated_FGFR3_G382D_mutant_(complex)
FGFR3_cysteine_mutants_(family)
Activated_FGFR3_K650E_mutant_(complex)
Activated_FGFR3_K650T_mutant_(complex)
FGFR3_A319E_mutant_dimer_(complex)
FGFR3_K650T_mutant_dimer_(complex)
Activated_FGFR3_K650M_mutant_(complex)
Activated_FGFR3_795fs_139STOP_mutant_(complex)
Activated_FGFR3c_homodimer_bound_to_FGF_(complex)
Activated_FGFR3_A391E_mutant_(complex)
Activated_FGFR3_G380R_mutant_(complex)
Activated_FGFR3b_G967C_(complex)
Activated_FGFR3_mutants_with_enhanced_kinase_activity_(family)
Activated_FGFR3_K650Q_mutant_(complex)
Activated_FGFR3c_P250R_mutant_(complex)
Activated_FGFR3b_S249C_mutant_(complex)
FGFR3c_homodimer_bound_to_FGF_(complex)
FGFR3c-binding_FGFs_(family)
FGFR3-2
FGFR3b-binding_FGFs_(family)
Activated_FGFR3b_homodimer_(complex)
Activated_FGFR3b_homodimer_bound_to_FGF_(complex)
FGFR3_mutant_dimers_TKIs_(complex)
FGFR3b_homodimer_(complex)
FGFR3b_homodimer_bound_to_FGF_(complex)
FGFR2c_short_homodimer_bound_to_FGF_(complex)
Activated_FGFR2c_short_homodimer_bound_to_FGF_(complex)
FGFR4/FGF19_(complex)
FGFR2c-binding_FGFs_(family)
FGF19
Activated_FGFR4_homodimer_(complex)
FGF18
FGF17
FGFR4-binding_FGFs_(family)
FGFR4_homodimer_bound_to_FGF_(complex)
FGF9
FGFR1c_homodimer_bound_to_FGF_(complex)
FGFR1c-binding_FGFs_(family)
Activated_FGFR1c_homodimer_(complex)
FGFR1_P252X_mutant_dimers_bound_to_FGFs_(complex)
Activated_FGFR1_P252X_mutants_(complex)
b
● Adenocarcinoma (31)
● Squamous (144)
Spearman Rho = −0.30
p=6.4E−5
FGFR3 mRNA Expression Levels (Log2(RSEM + 1))
4
6
8
10
12
14
6
8
10
12
14
FGFR1 mRNA Expression Levels (Log2(RSEM + 1))
c
FGFR1 mRNA
FGFR3 mRNA
Log2(RSEM + 1)Log2(RSEM + 1)
4
6
8
10
12
14
Adeno (31)
Squam (144)
Adeno (31)
Squam (144)

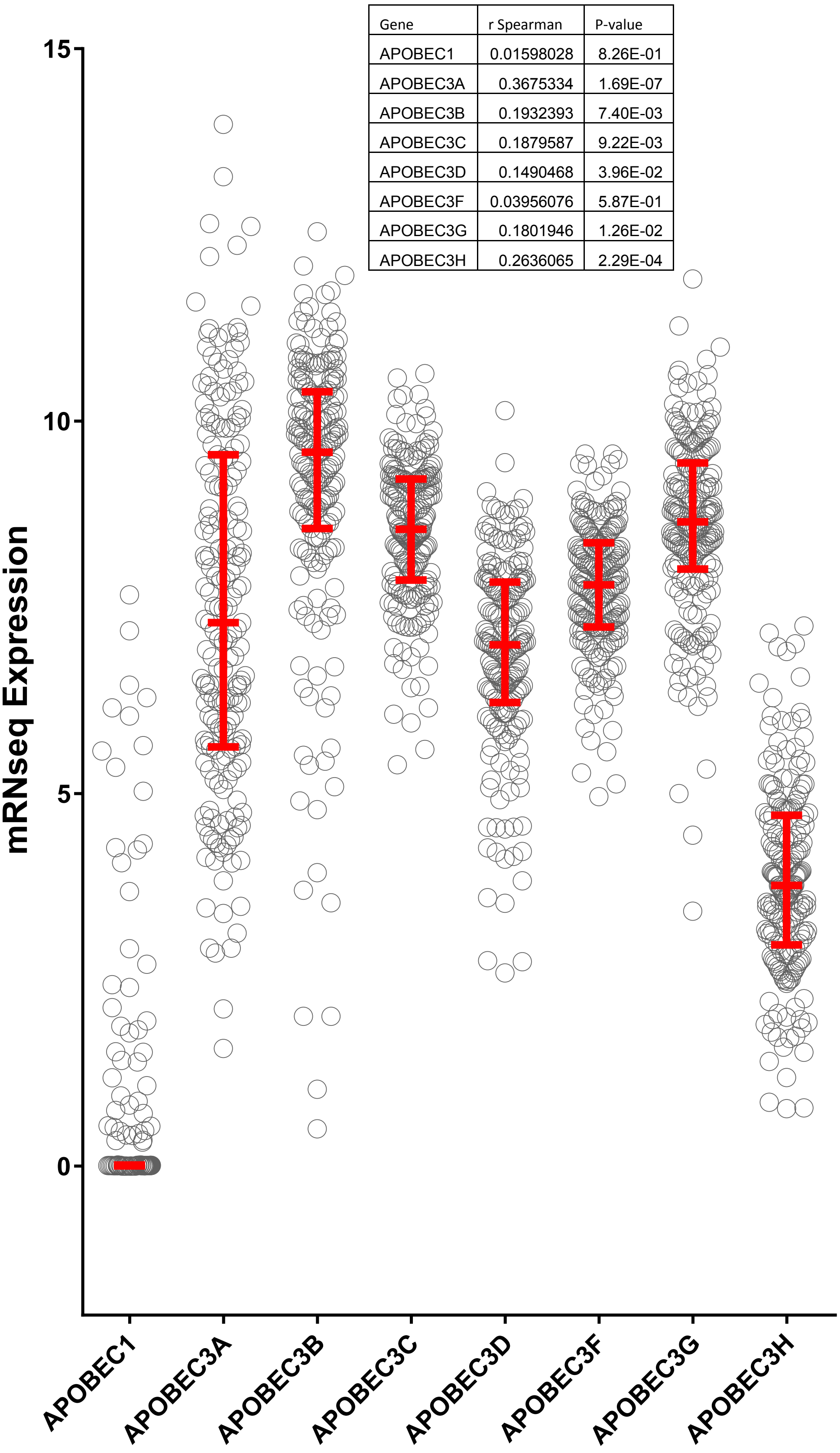
Supplemental Figure S25: FGF signaling in squamous carcinomas relative to adenocarcinomas. **a,** Cytoscape view of interconnected PARADIGM features differentially activated between squamous carcinoma and adenocarcinoma. Node color reflects level of differential activation (red: higher in squamous, blue: higher in adenocarcinoma), while node size reflects significance. Edge color and arrow denotes interaction type (purple arrow: activation, green T: inhibition, black •: component link). Genes with differential PARADIGM inferred activities are highlighted in bold text. **b,** Scatterplot of FGFR3 vs. FGFR1 mRNA expression levels. **c,** Boxplots of FGFR1 and FGFR3 mRNA expression in squamous (Squam) and adenocarcinoma (Adeno) samples. Colors reflect histology (cyan: squamous, gray: adenocarcinoma).



Supplemental Figure S26: A high level of APOBEC mutagenesis pattern is present in many TCGA cervical cancer samples. **a,** Fractions of total mutation counts in a sample. **b,** Minimum estimate of the number of APOBEC-induced mutations in a sample (the log10 scale with a pseudo count of 0.1 is used to show samples with "APOBEC_MutLoad_MinEstimate"=0).



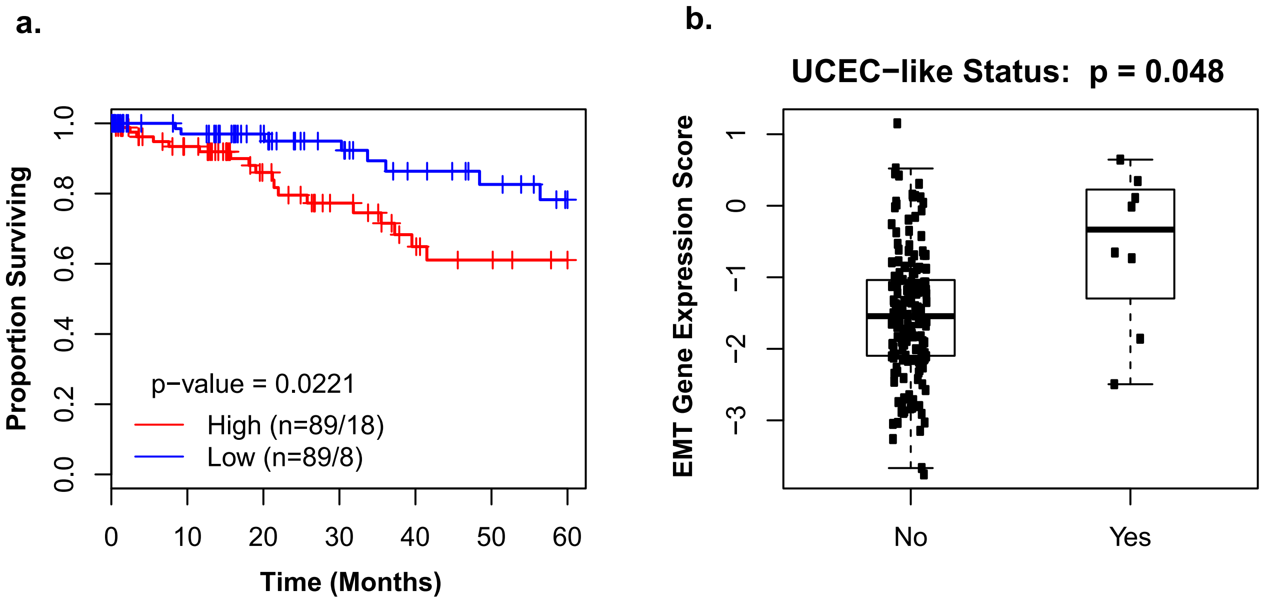
Supplemental Figure S27: APOBEC mutagenesis pattern is reduced in HPV-negative samples compared with HPV-positive samples. APOBEC mutagenesis patterns are presented for samples of different HPV clade (HPV A7, HPV A9, and HPV Other) and status (HPV-negative and HPV-positive).



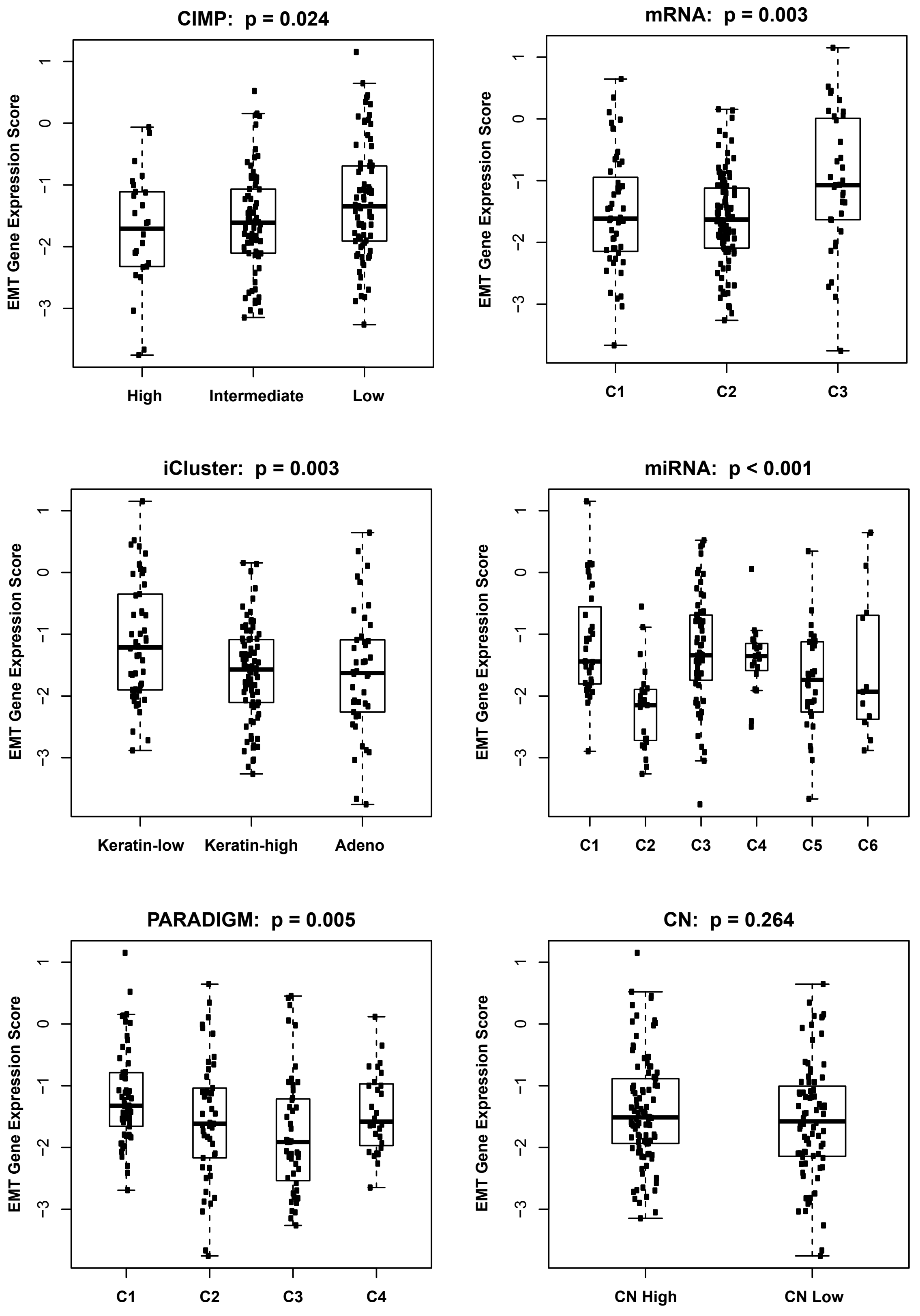
Gene	r Spearman	P-value
APOBEC1	0.01598028	8.26E-01
APOBEC3A	0.3675334	1.69E-07
APOBEC3B	0.1932393	7.40E-03
APOBEC3C	0.1879587	9.22E-03
APOBEC3D	0.1490468	3.96E-02
APOBEC3F	0.03956076	5.87E-01
APOBEC3G	0.1801946	1.26E-02
APOBEC3H	0.2636065	2.29E-04

Supplemental Fig. S28

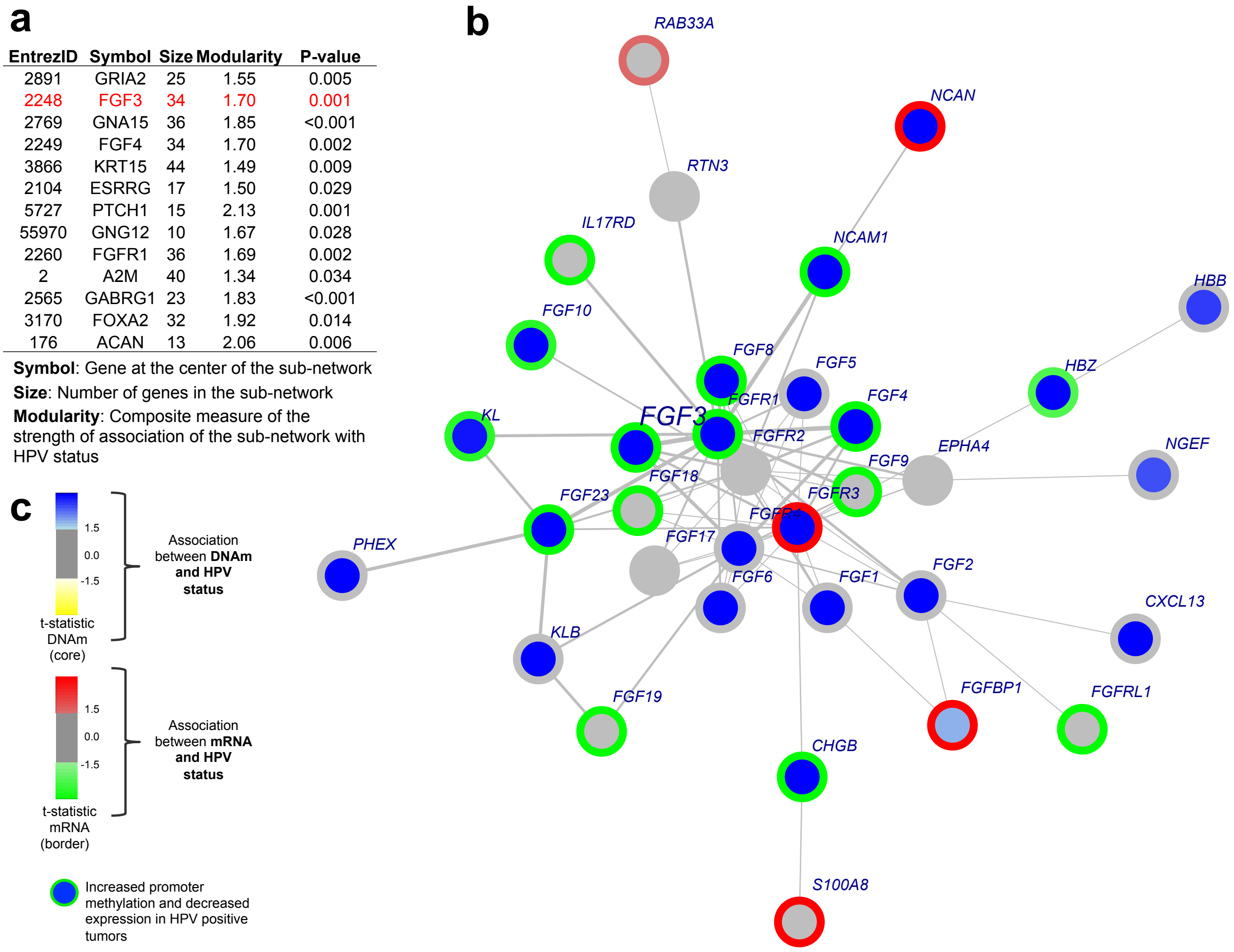
Supplemental Fig. S28: Expression of *APOBEC* genes in TCGA cervical cancer samples correlates with APOBEC mutagenesis. mRNA gene expression (log2(RSEM+1)) is plotted for *APOBEC1* and *APOBEC3* genes across all 192 Extended Set samples. A small number of zero values for *APOBEC1* expression are not plotted, but the complete set of values were used to calculate r-Spearman correlation values between expression and "APOBEC_MutLoad_MinEstimate" shown in the table-insert.



Supplemental Figure S29: EMT mRNA score associations with outcomes and UCEC-like status. **a**, Five year survival analysis comparing the EMT-high vs. EMT-low groups (log-rank p =0.0221). **b**, Association of EMT scores with UCEC-like case status (two sample t-test, p=0.048).



Supplemental Figure S30: Association of EMT mRNA scores with platform clusters. Association of EMT scores with methylation CIMP, iCluster, miRNA, mRNA, PARADIGM, and copy number (CN) clusters using AVONA test is presented.



a

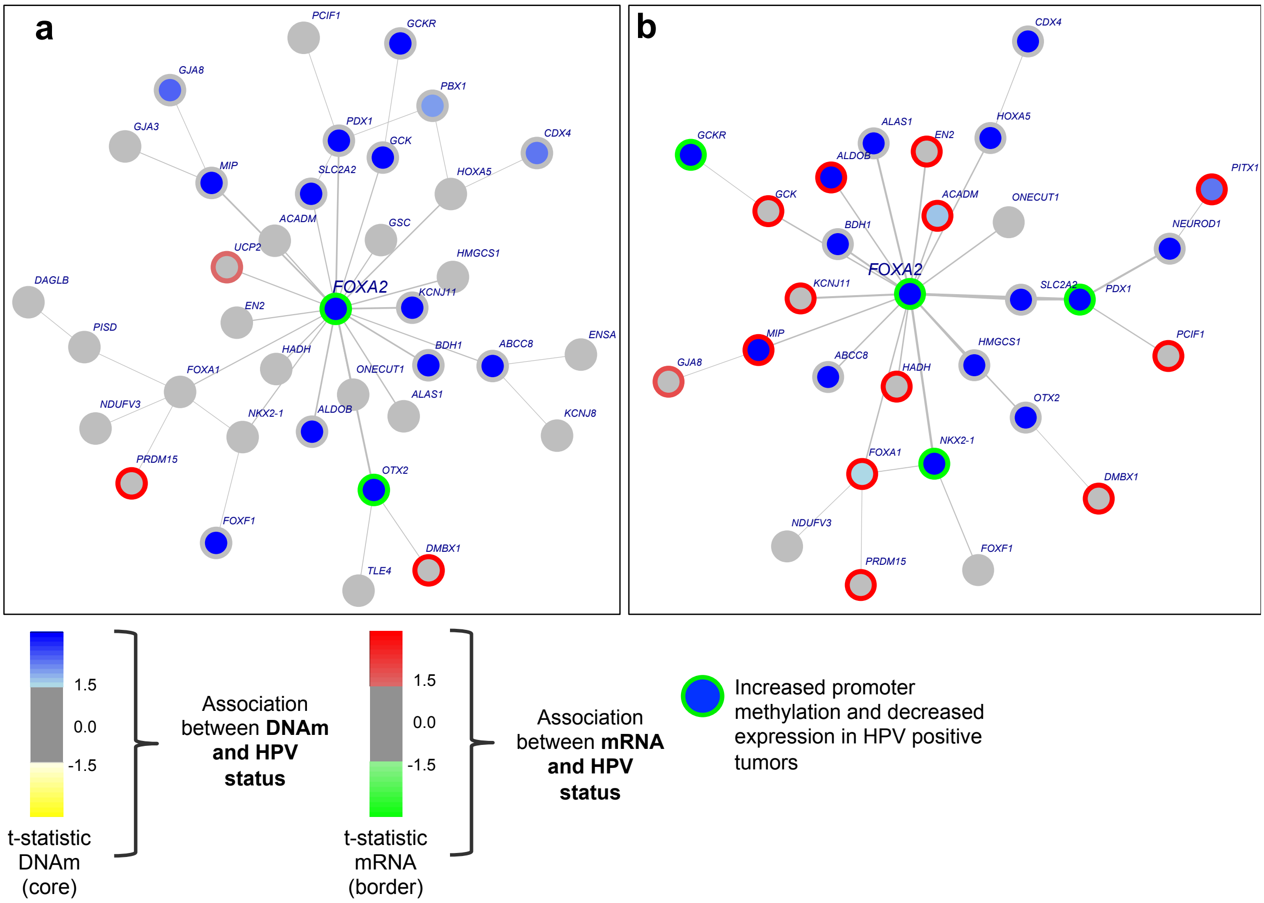
EntrezID	Symbol	Size	Modularity	P-value
2891	GRIA2	25	1.55	0.005
2248	FGF3	34	1.70	0.001
2769	GNA15	36	1.85	<0.001
2249	FGF4	34	1.70	0.002
3866	KRT15	44	1.49	0.009
2104	ESRRG	17	1.50	0.029
5727	PTCH1	15	2.13	0.001
55970	GNG12	10	1.67	0.028
2260	FGFR1	36	1.69	0.002
2	A2M	40	1.34	0.034
2565	GABRG1	23	1.83	<0.001
3170	FOXA2	32	1.92	0.014
176	ACAN	13	2.06	0.006

Symbol: Gene at the center of the sub-network

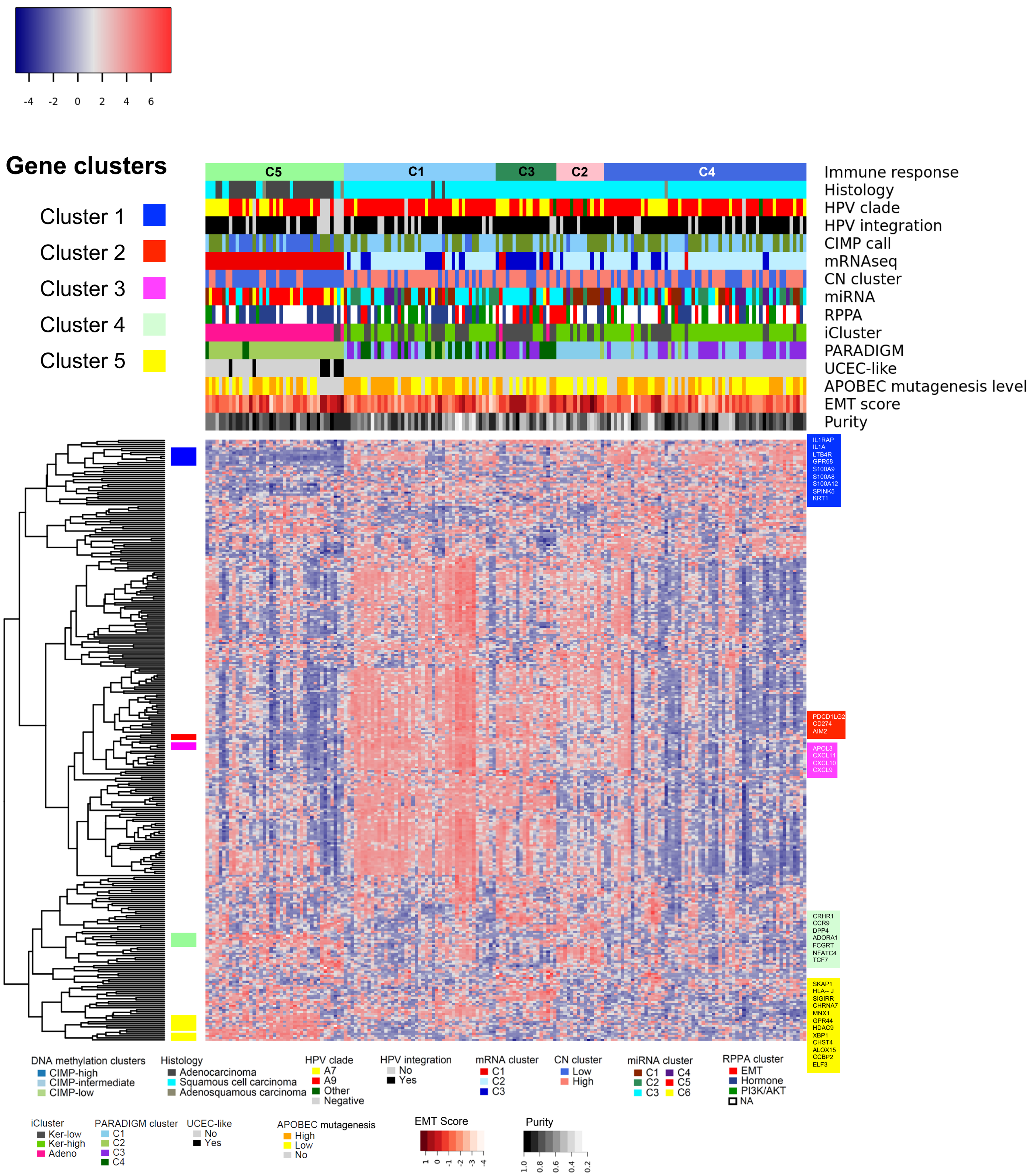
Size: Number of genes in the sub-network

Modularity: Composite measure of the strength of association of the sub-network with HPV status

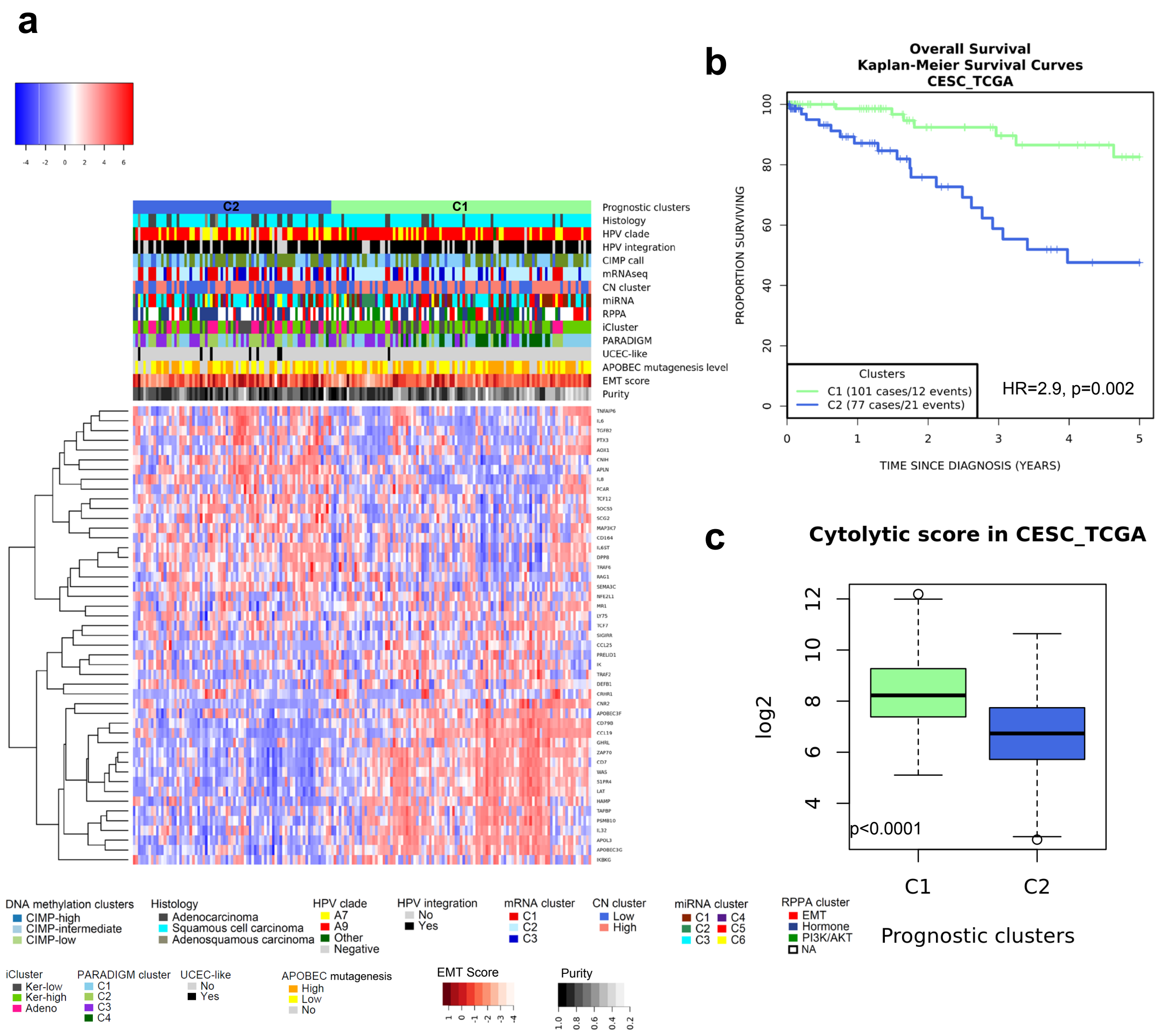
Supplemental Figure S31: Statistically significant results from the Functional Epigenetic Module (FEM) analysis for identifying disrupted subnetworks between HPV-negative (n = 9) and HPV-positive (n = 169) cervical tumors. **a,** Table of the 13 statistically significant (p < 0.05) subnetworks. Genes contained in this table refer to central gene (node) of the subnetwork. Size refers to the number of genes in a particular subnetwork. Modularity is a composite measure of the strength of association between HPV status and expression/DNA methylation (DNAm) of the genes within a subnetwork. **b,** Subnetwork centered around Fibroblast Growth Factor 3 (*FGF3*) consisting of 34 genes. **c,** Description of the color scheme for node cores and borders.



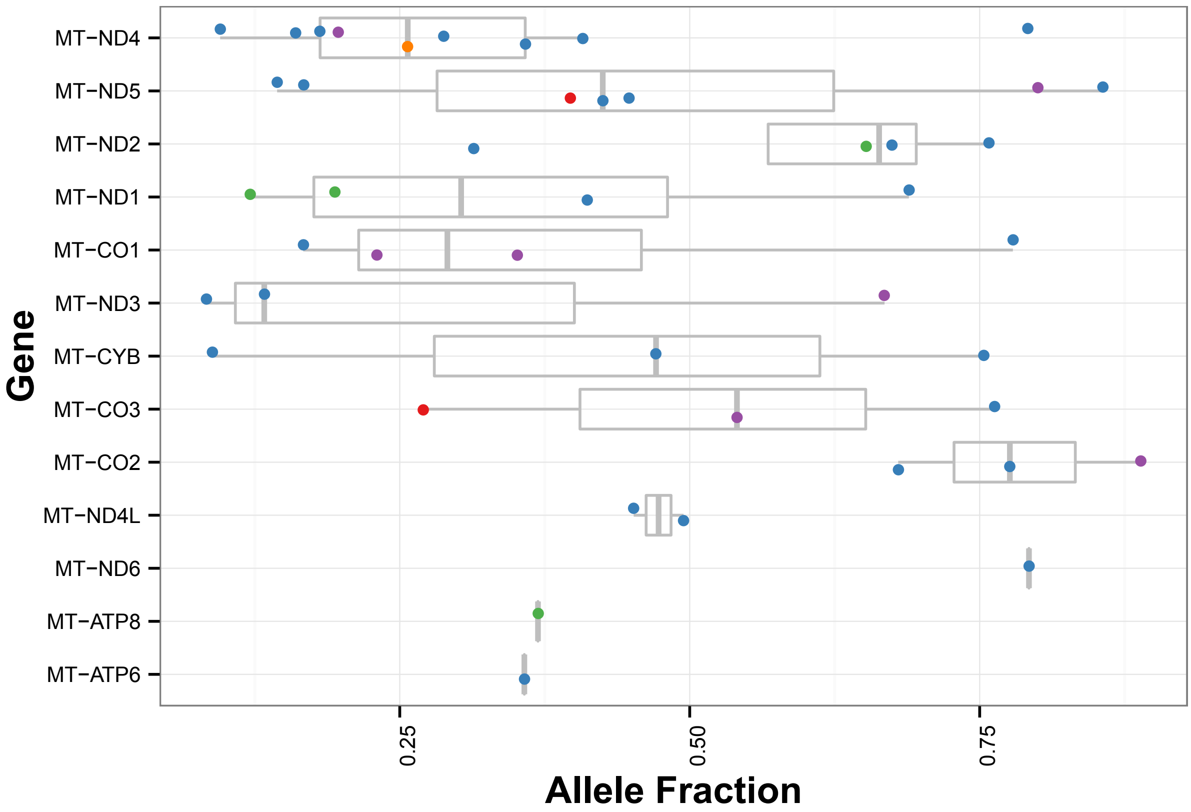
Supplemental Figure S32: FEM analysis of cervical and head and neck squamous cell carcinomas. Significantly disregulated subnetwork centered around Forkhead Box A2 (*FOXA2*) between HPV-positive and -negative cervical (**a**) and head and neck (**b**) squamous cell carcinomas.



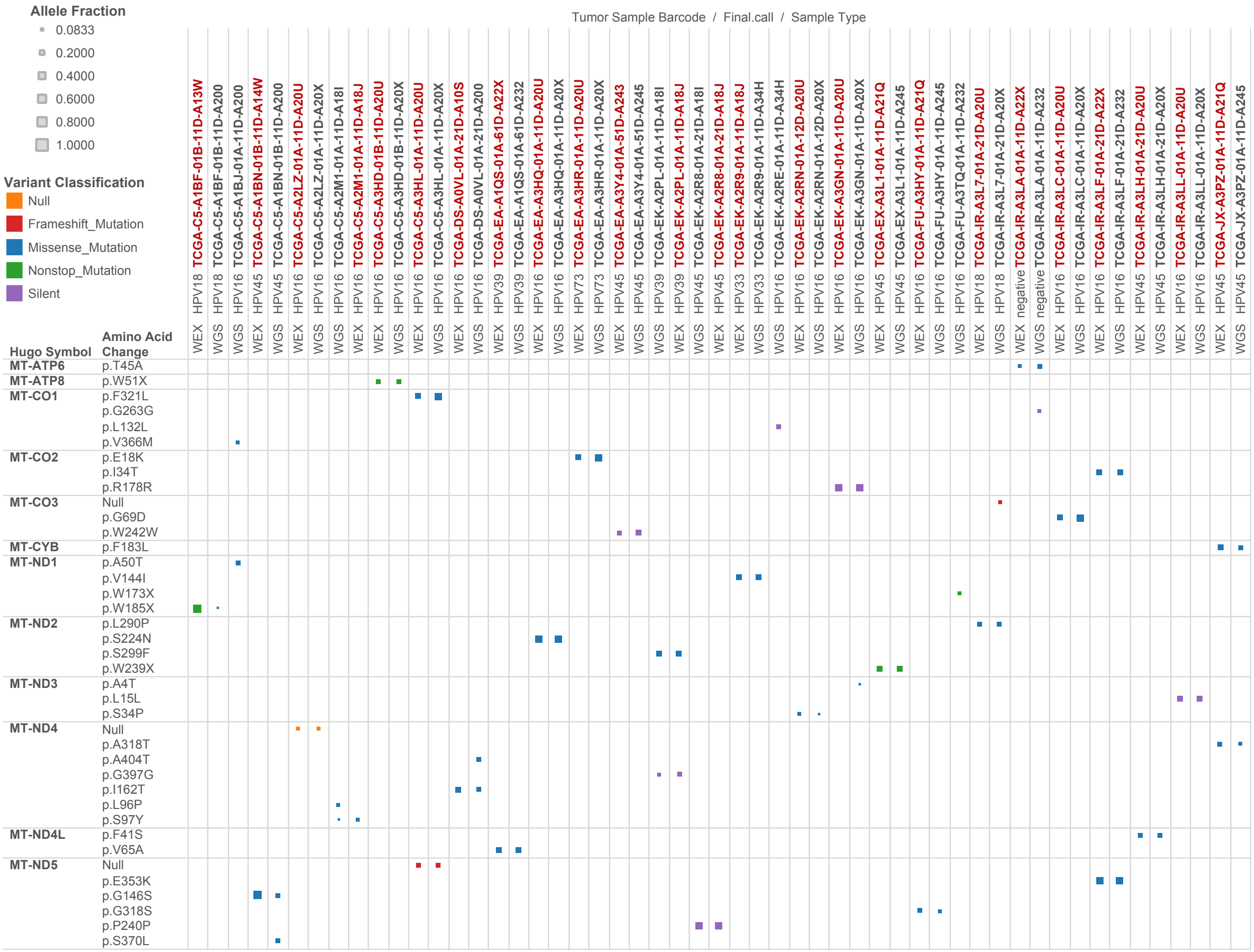
Supplemental Figure S33: **Immune response gene clustering analysis**. Consensus clustering analysis of 178 cervical cancer samples based on immune response gene expression. Gene clusters represent significant genes comparing C5 samples with all other samples.



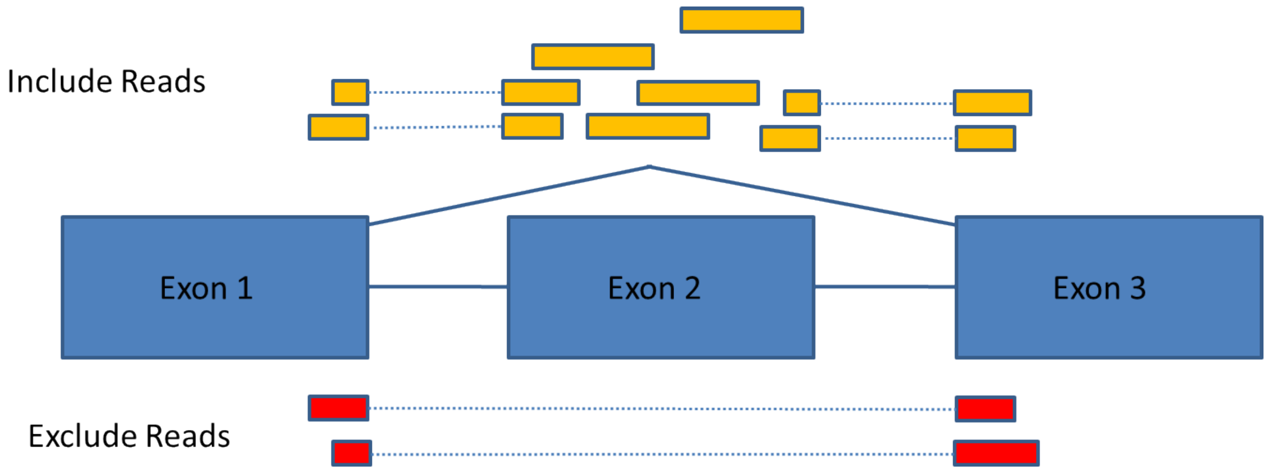
Supplemental Figure S34: **Immune response prognostic gene clustering analysis**. **a,** Consensus clustering analysis of 178 cervical carcinoma samples based on immune response gene expression selected from ROC analysis. **b,** Kaplan-Meier analysis comparing survival across predicted prognostic clusters. **c,** Comparison of cytolytic scores across prognostic clusters.



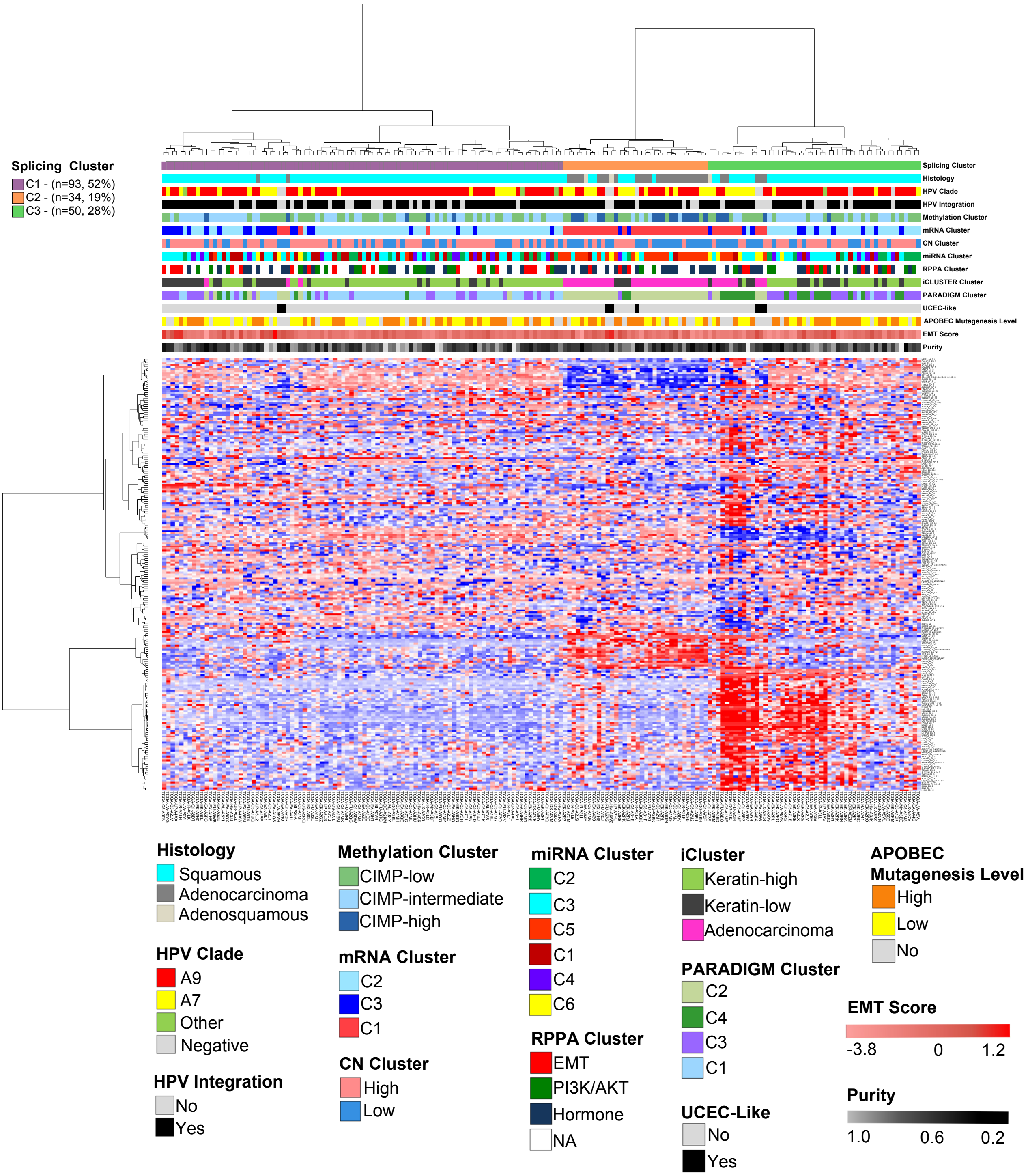
Supplemental Figure S35: Somatic mitochondrial gene mutations in cervical cancer. A total of 45 mitochondrial mutations were found across 31 of 50 samples analyzed using WGS (5-7X). Each point represents a mutation detected at the given allele fraction, with the color indicating the variant classification (orange, Null; red, Frameshift mutation; blue, Missense mutation; green, Nonsense mutation; purple, Silent mutation). Among the 13 mitochondrial genes, *ND4* had most (9) mutations. Three out of 50 cases had a mutation in *CO2* gene, all with a high allelic fraction.



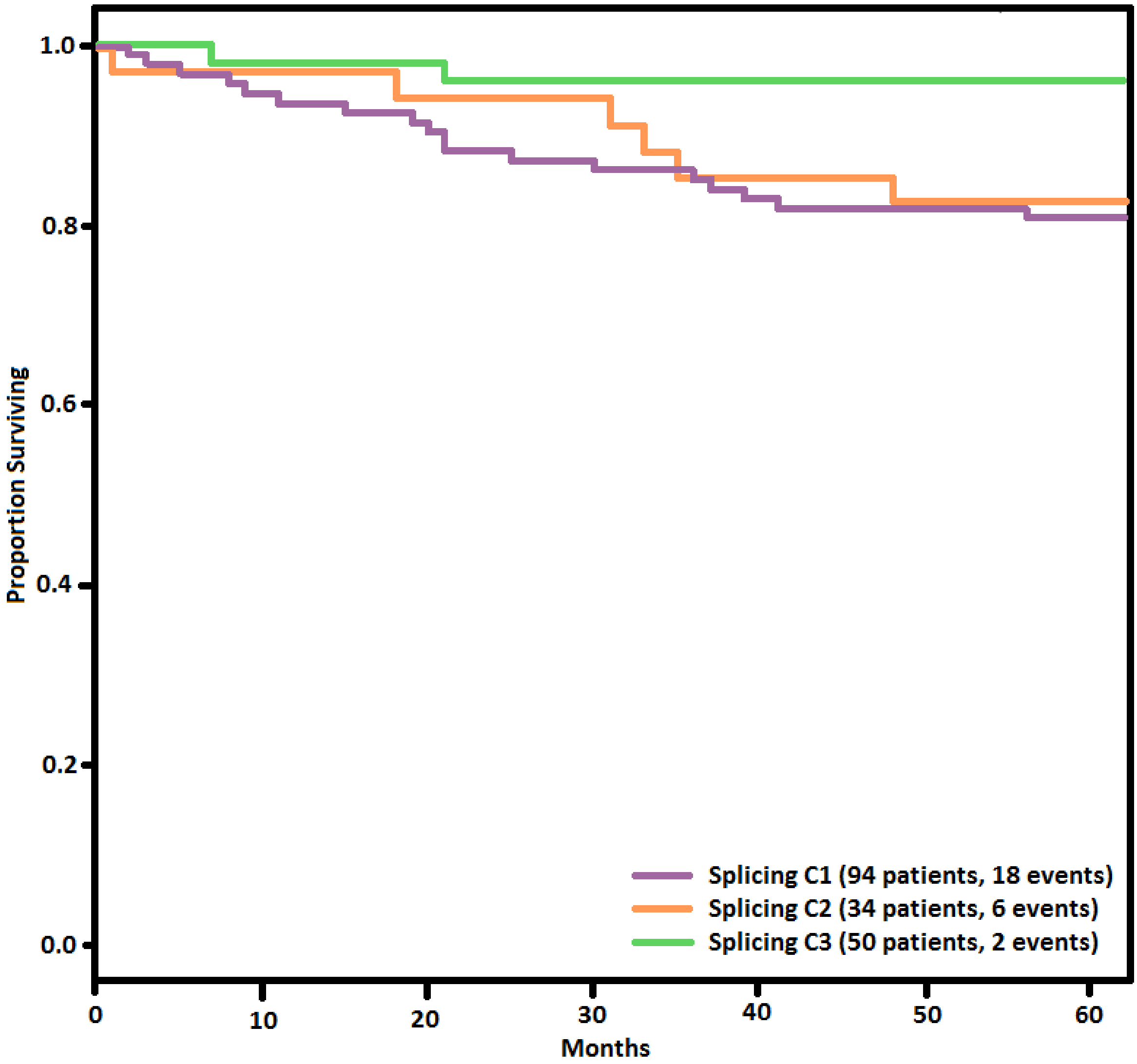
Supplemental Figure S36: Somatic mitochondrial gene mutations in cervical cancer by patient. Specific mutations are listed for each patient tumor and color coded based on mutation type for whole genome (WGS; black) or whole exome (WEX; red) sequencing. The size of each box represents the magnitude of allele fraction.



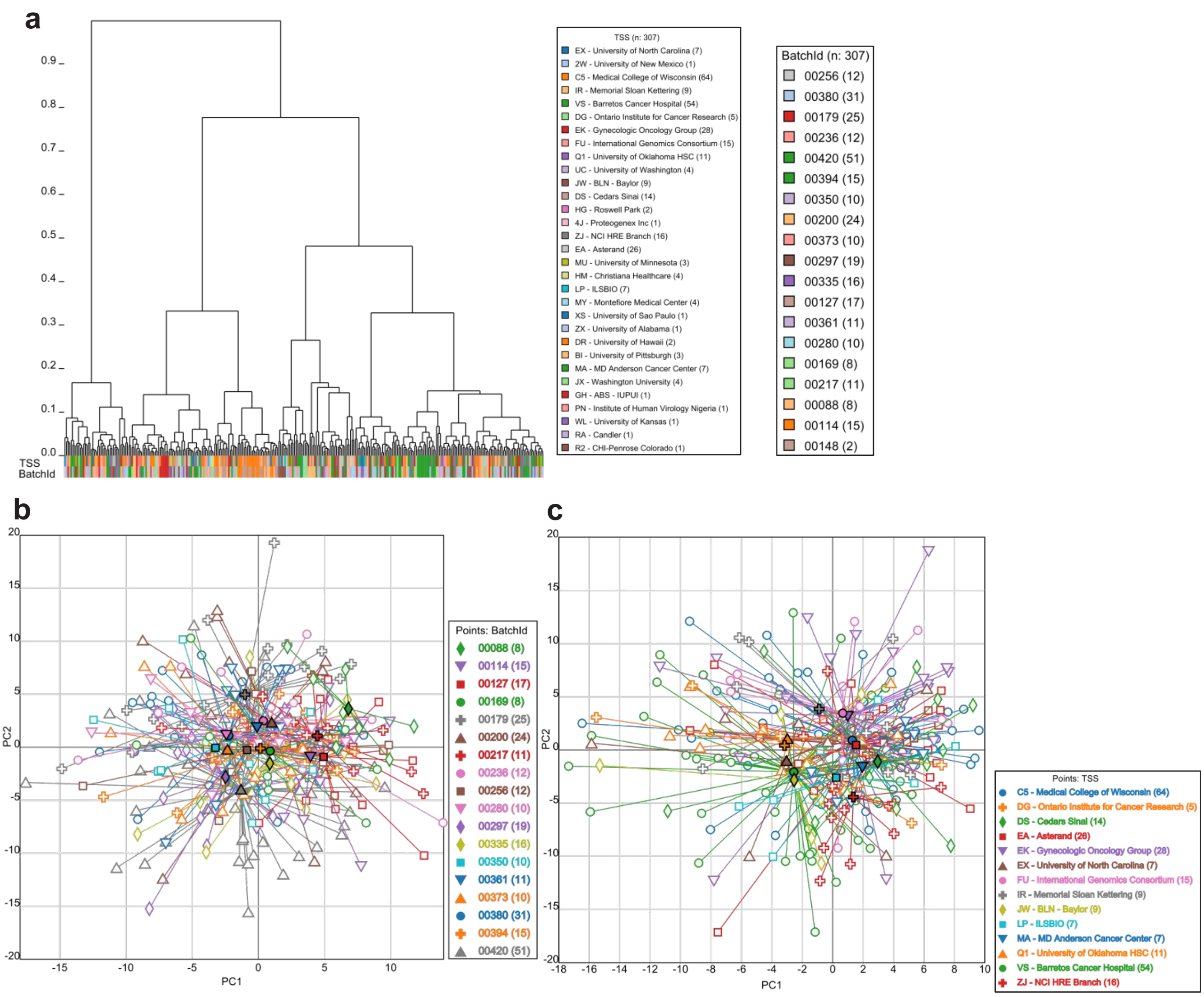
Supplemental Figure S37: PSI as determined from an exon skipping event. The yellow exon and exon junction reads indicate the presence of exon 2 while the red exon 1 – 3 junction reads indicate that exon 2 is spliced out. The PSI is 8 / 10 reads or .8 indicating that ~ 80% of transcripts in the sample include exon 2 and 20% do not.



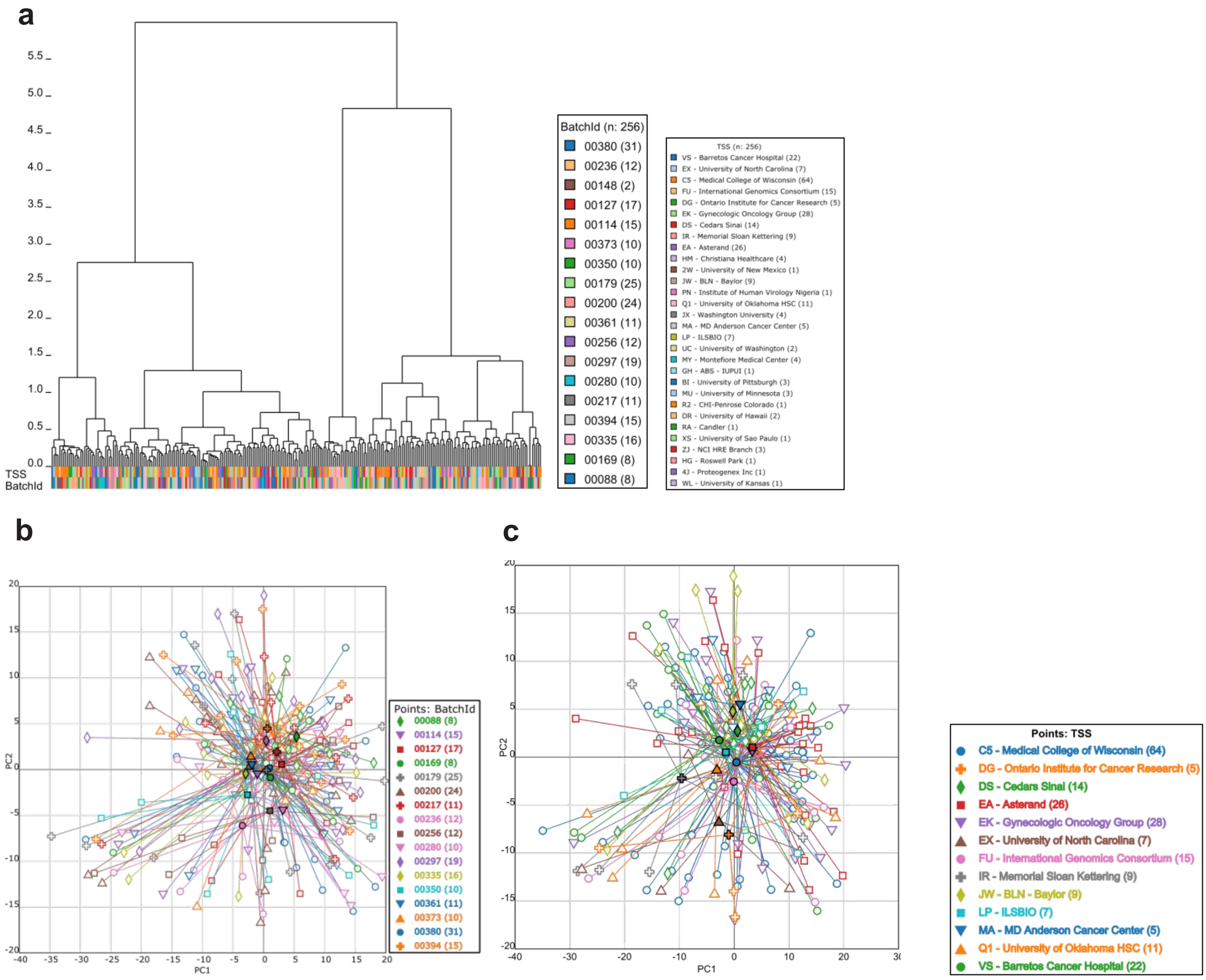
Supplemental Figure S38: Hierarchal clustering heatmap based on splicing events. A hierarchical clustering heatmap is shown of mean-centered PSI values for high variation splice events in samples.



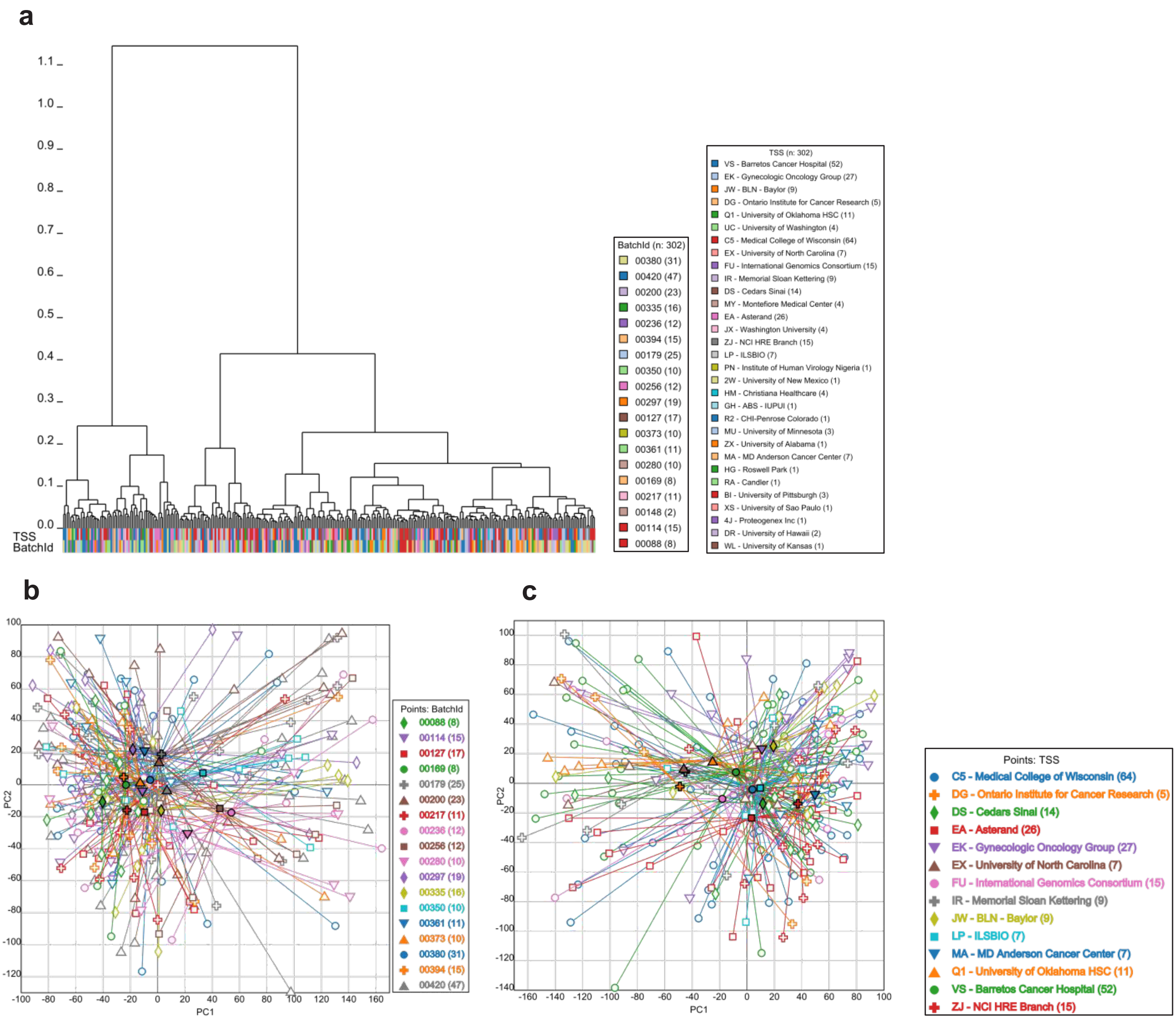
Supplemental Figure S39: Survival analysis for the three splicing clusters. Kaplan-Meier analysis was performed to assess survival differences across all three splicing clusters.



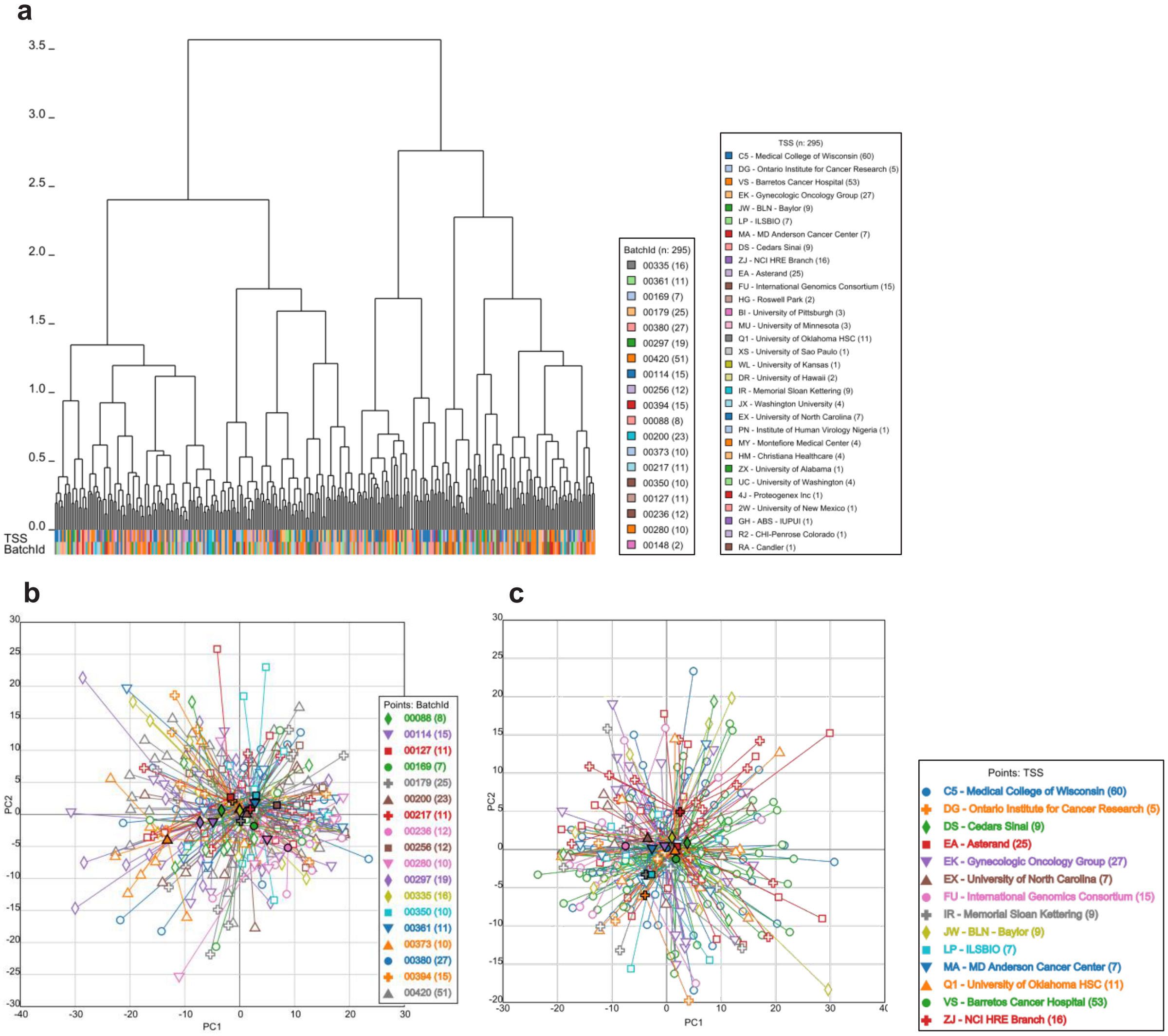
Supplemental Figure S40: miRNA batch effects. **a**, Hierarchical clustering of samples based on miRNA sequencing data. **b**, PCA for miRNA expression from miRNAseq data, with samples connected by centroids according to batch ID. **c**, PCA for miRNA expression from miRNAseq data, with samples connected by centroids according to TSS.



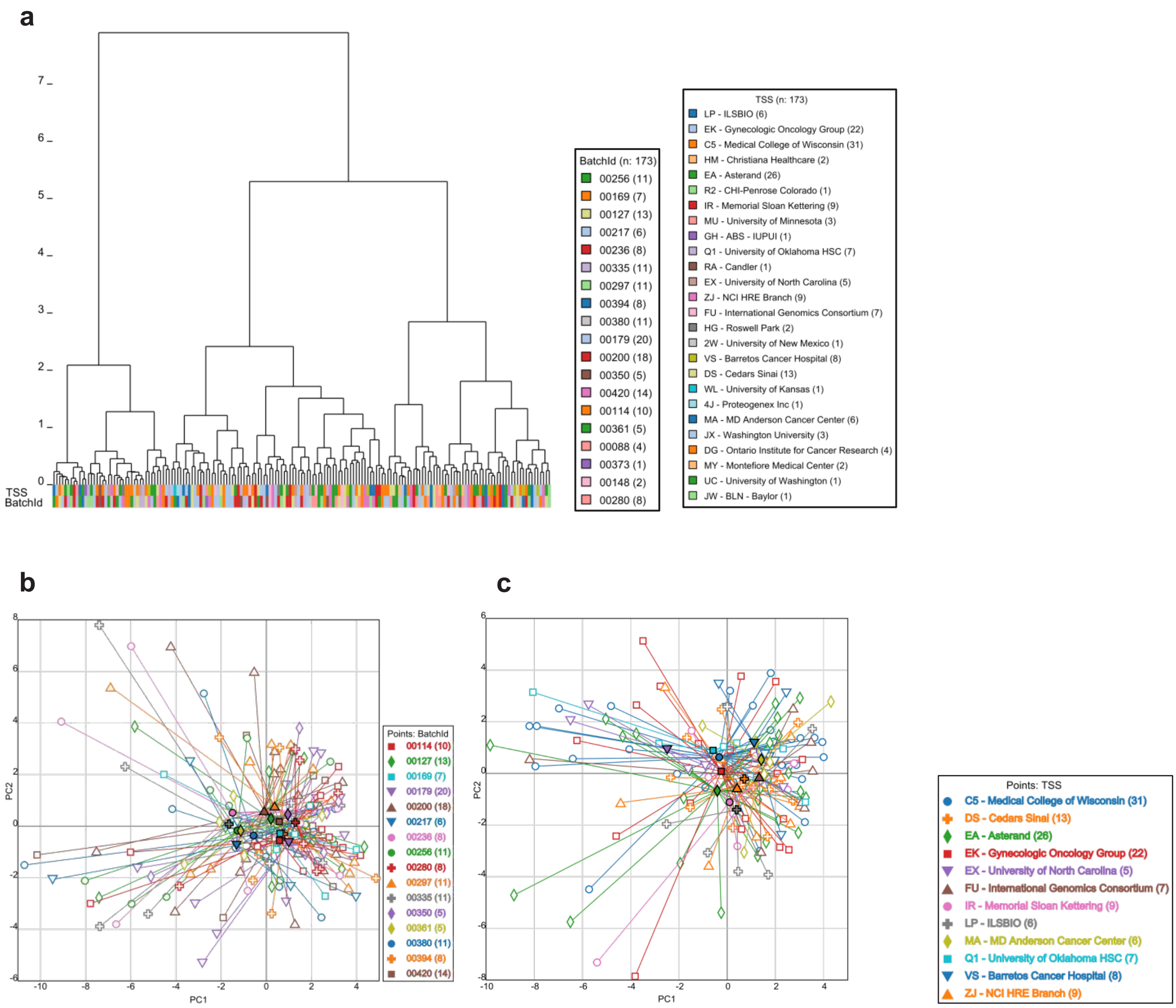
Supplemental Figure S41: DNA methylation batch effects. **a**, Hierarchical clustering of samples based on DNA methylation data. **b**, PCA for DNA methylation, with samples connected by centroids according to batch ID. **c**, PCA for DNA methylation, with samples connected by centroids according to TSS.



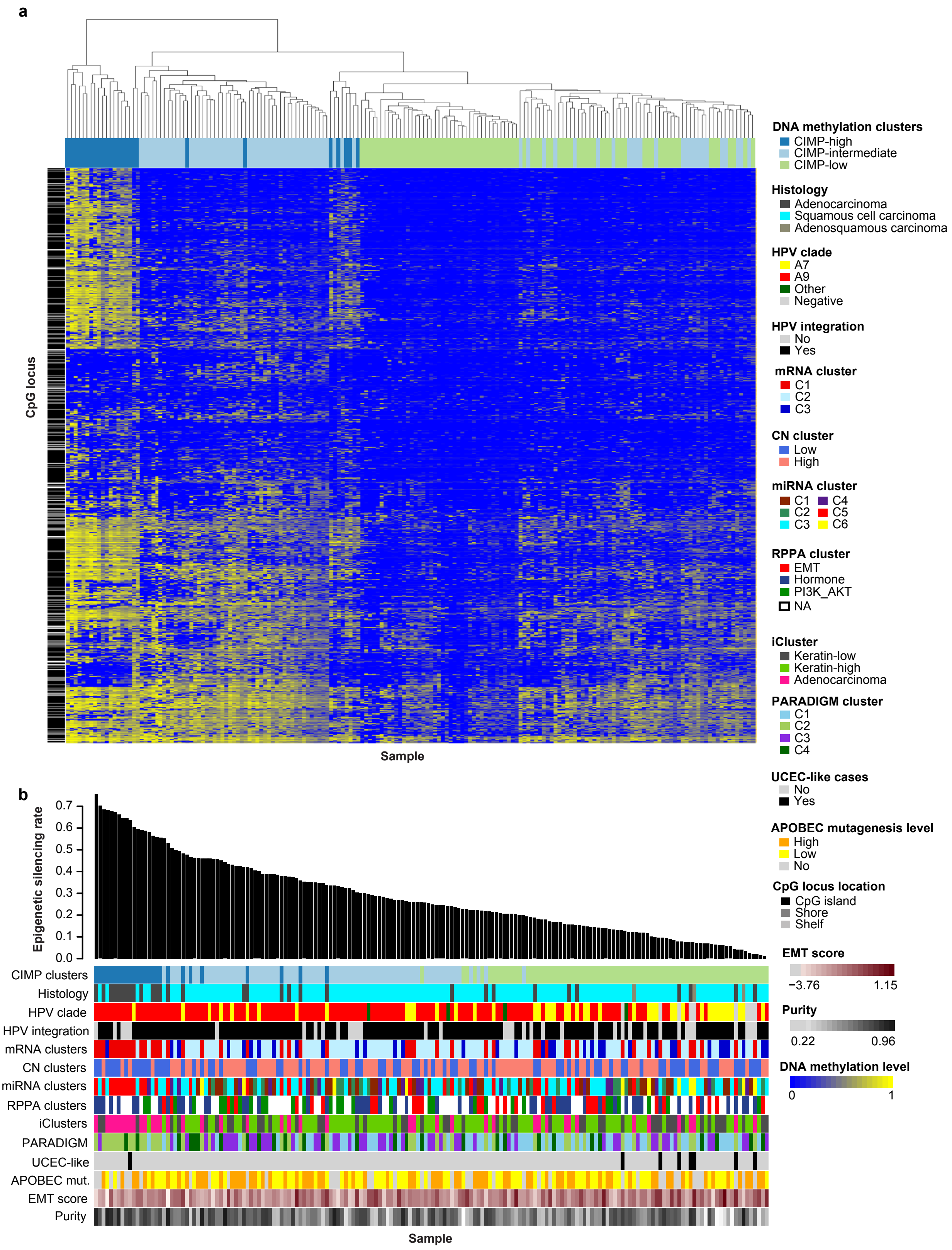
Supplemental Figure S42: RNAseq batch effects. **a**, Hierarchical clustering of samples based on RNAseq expression data. **b**, PCA for RNAseq, with samples connected by centroids according to batch ID. **c**, PCA for RNAseq, with samples connected by centroids according to TSS.



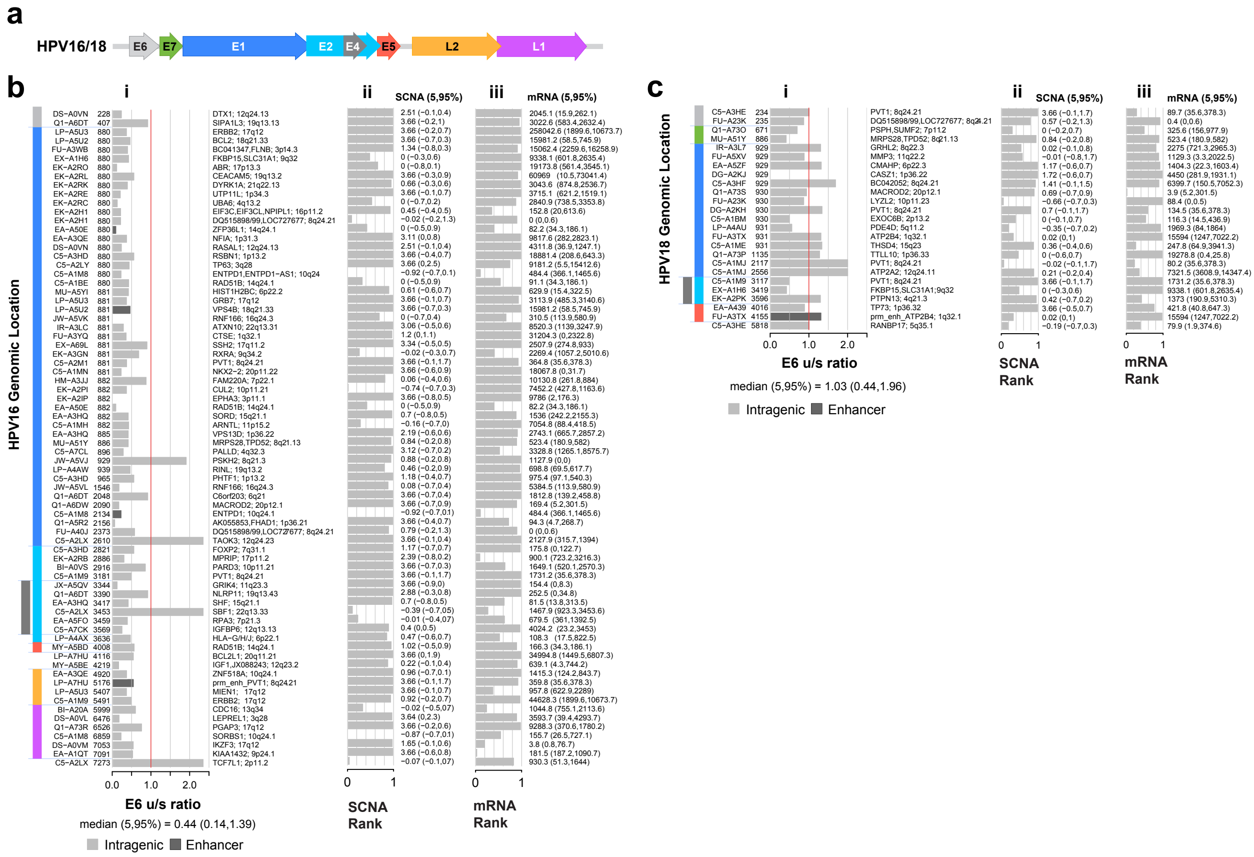
Supplemental Figure S43: SNP6 copy number batch effects. **a**, Hierarchical clustering of samples based on SNP6 copy number data. **b**, PCA for SNP6 data, with samples connected by centroids according to batch ID. **c**, PCA for SNP6 data, with samples connected by centroids according to TSS.



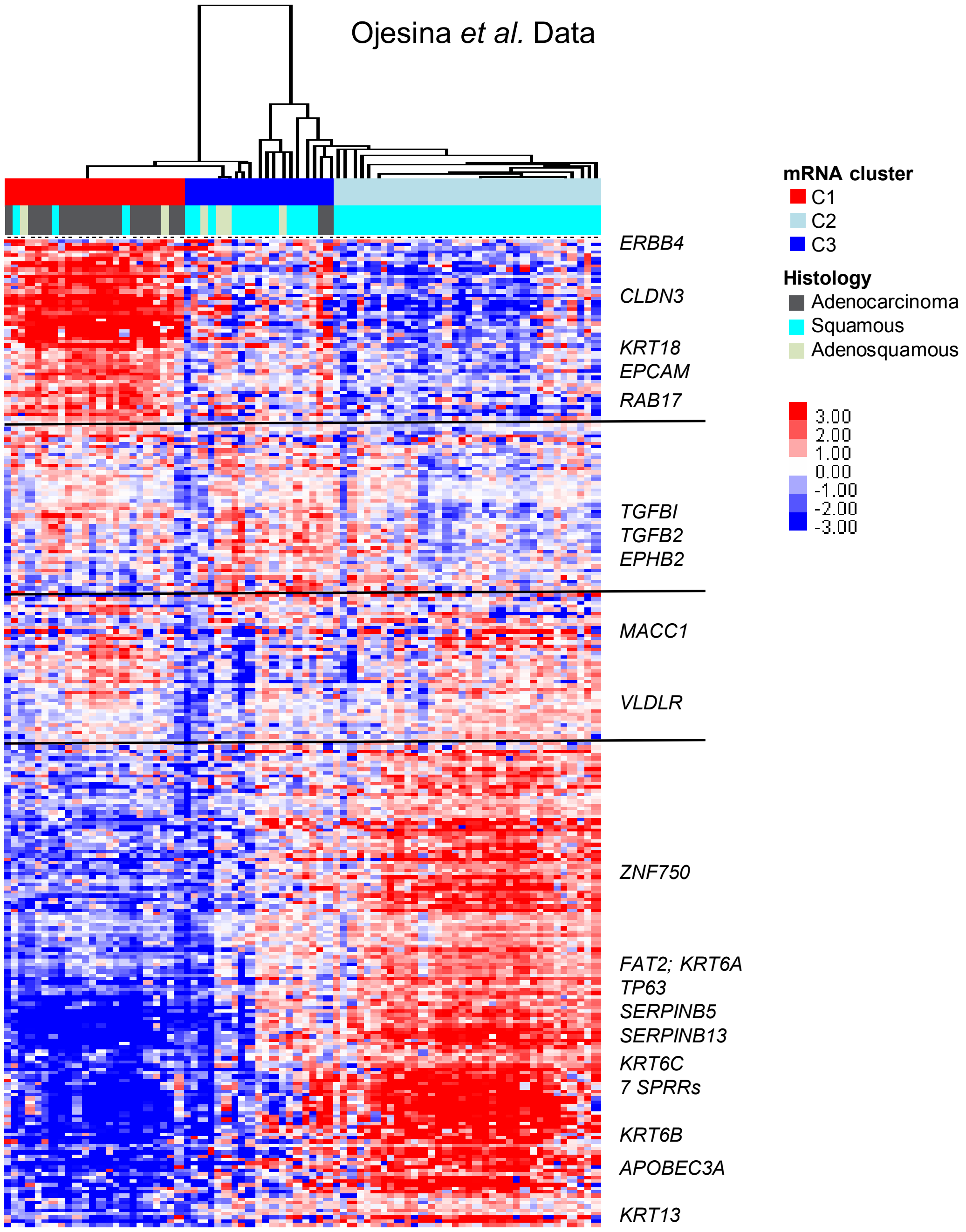
Supplemental Figure S44: RPPA batch effects. **a**, Hierarchical clustering of samples based on RPPA data. **b**, PCA for RPPA data, with samples connected by centroids according to batch ID. **c**, PCA for RPPA data, with samples connected by centroids according to TSS.


a
CpG locus
Sample
DNA methylation clusters
CIMP-high
CIMP-intermediate
CIMP-low
Histology
Adenocarcinoma
Squamous cell carcinoma
Adenosquamous carcinoma
HPV clade
A7
A9
Other
Negative
HPV integration
No
Yes
mRNA cluster
C1
C2
C3
CN cluster
Low
High
miRNA cluster
C1
C4
C2
C5
C3
C6
RPPA cluster
EMT
Hormone
PI3K_AKT
NA
iCluster
Keratin-low
Keratin-high
Adenocarcinoma
PARADIGM cluster
C1
C2
C3
C4
UCEC-like cases
No
Yes
APOBEC mutagenesis level
High
Low
No
CpG locus location
CpG island
Shore
Shelf
EMT score
−3.76
1.15
Purity
0.22
0.96
DNA methylation level
0
1
b
Epigenetic silencing rate
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
CIMP clusters
Histology
HPV clade
HPV integration
mRNA clusters
CN clusters
miRNA clusters
RPPA clusters
iClusters
PARADIGM
UCEC-like
APOBEC mut.
EMT score
Purity
Sample

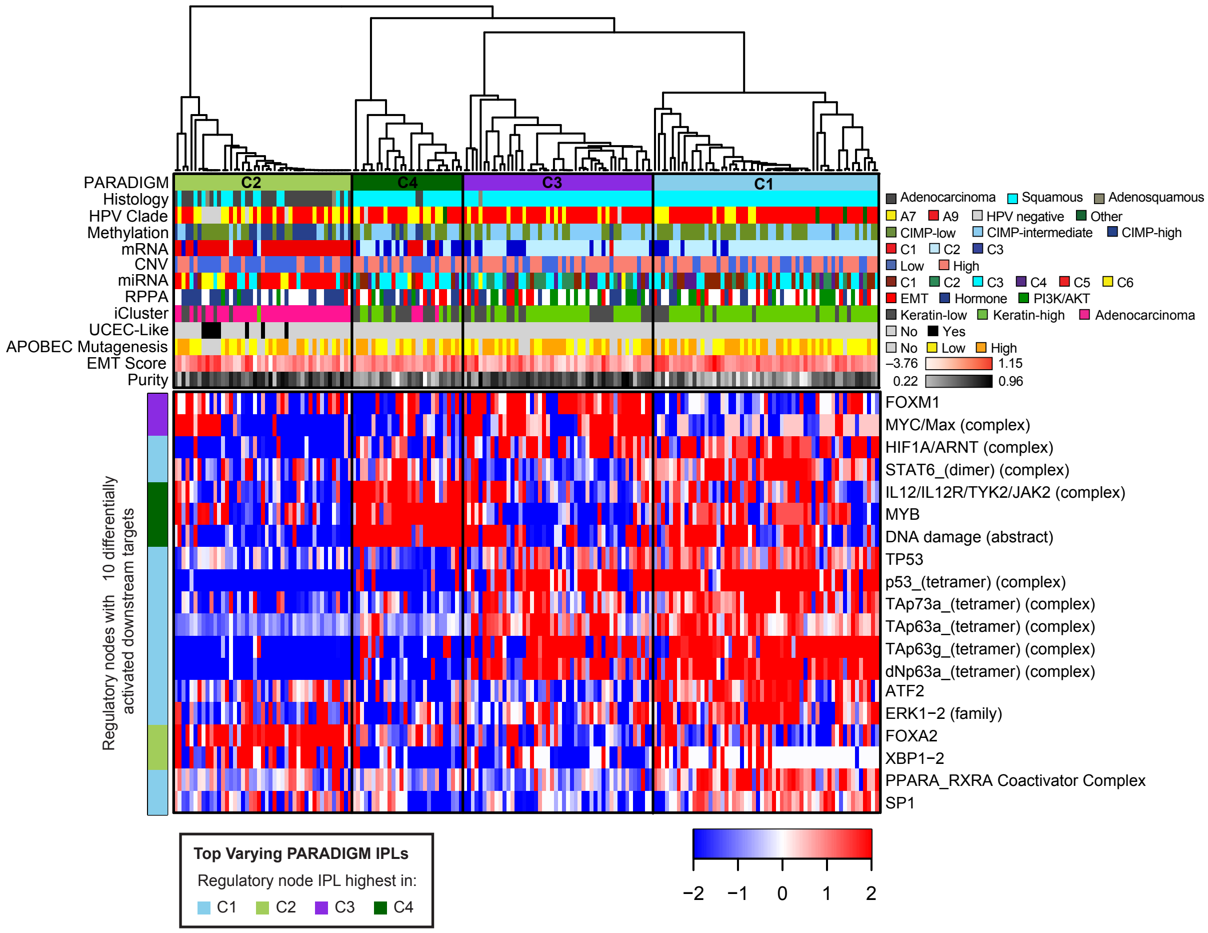
Supplemental Figure S45: Epigenetically silenced gene frequencies. **a**, Heatmap showing unsupervised hierarchical clustering of beta values of 178 samples and epigenetically silenced probes from Supplemental Table 12. Samples are presented in columns and CpG loci are presented in rows. An annotation panel on the top of the heatmap indicates CIMP clusters. An annotation panel on the left of the heatmap indicates CpG locus location. **b**, Ordered sample-wise epigenetic silencing rates and corresponding annotation panel, which shows that samples with higher silencing rates were HPV clade A9-positive adenocarcinomas.



Supplemental Figure S46: HPV integration sites from RNAseq data. Sites are alignment locations for viral-human junctions in chimeric contigs generated by *de novo* assembly of RNAseq data. **a**, A schematic genome of HPV16 and HPV18. Note: HPV16 and HPV18 reading frames differ for E2/E4 and L1/L2. **b-c**, HPV16 (b) and HPV18 (c) integration into genes and enhancers. i, Left to right: genomic locations (from PAVE HPV16REF.1, GI:333031 and HPV18REF.1, GI:60975; pave.niaid.nih.gov), sample ID, HPV coordinate for the viral-human junction in a chimeric RNAseq contig from *de novo* assembly, E6 unspliced/spliced ratio (E6 u/s ratio), gene (or enhancer-associated gene) and cytoband into which the viral-human junction in a chimeric RNAseq contig aligns. ii,iii, Ranks of the somatic copy number alteration (SCNA) and the mRNA abundance for human genes in (i). Ranks indicate the SCNA or mRNA value for the sample and gene with the chimeric junction relative to values for that gene across all samples, with 0 and 1 representing the lowest and highest ranked values, respectively. Bars show the rank of the integrated gene within the specified sample, while text presents the actual SCNA or mRNA abundance value, with 5th and 95th percentile values.



Supplemental Figure S47: mRNA clustering of Ojesina *et al.* dataset. Hierarchical clustering of mRNA data from 75 cervical cancer samples from Ojesina *et al.* (Ojesina, A.I. *et al.* Landscape of genomic alterations in cervical carcinomas. Nature. 506, 371-375 (2014)). Four samples were dropped because they were of histological subtypes that were not represented in the TCGA set.



Supplemental Figure S48: PARADIGM pathway activity clusters. Consensus clustering of top 25% of varying PARADIGM integrated pathway level features (IPLs) was performed. PARADIGM IPLs were determined by integration of copy number, mRNA gene expression, and pathway interaction data. Samples are ordered by their consensus cluster membership. Histology, HPV clade, cluster assignments (methylation, mRNA, copy number (CNV), miRNA, RPPA, and iCluster), endometrial cancer-like (UCEC-like) annotation, APOBEC mutagenesis level, EMT mRNA score, and purity estimates for each sample are shown. The heatmap displays IPLs of regulatory hubs with at least 10 downstream targets showing differential inferred activation between PARADIGM clusters. Row annotation colors represent the PARADIGM cluster where the regulatory hub IPL is highest (C1: light blue C2: light green, C3: purple, C4: dark green). Heatmap scale represents median-centered IPL values.