

1 **Supplemental Information**

2
3 **Acknowledgments**
4

5 We would like to acknowledge the late Dr. Helga Salvesen (The University of Bergen), who
6 provided critical clinical and translational insight, and we dedicate this manuscript to her memory.

7 We also acknowledge Leslie Gaffney (The Broad Institute) for her work in preparing some of the
8 figures. In addition, this study was supported by National Institutes of Health (NIH) grants U54
9 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840,
10 U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24
11 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, and P30 CA016672.

12
13
14
15
16
17
18
19
20
21
22
23

24 **S1. Biospecimen Collection and Clinical Data**

25

26 **Sample Acquisition**

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

37

38 Pathology quality control was performed on each tumor and adjacent normal tissue (if available)
39 specimen from either a frozen section slide prepared by the Biospecimen Core Resource (BCR) or
40 from a frozen section slide prepared by the TSS. Hematoxylin and eosin (H&E) stained sections
41 from each sample were subjected to independent pathology review to confirm that the tumor
42 specimen was histologically consistent with the allowable cervical cancers and the adjacent normal
43 specimen contained no tumor cells. The percent tumor nuclei, percent necrosis, and other

44 pathology annotations were also assessed. Tumor samples with $\geq 60\%$ tumor nuclei and $\leq 20\%$
45 necrosis were submitted for nucleic acid extraction.

46

47 Approximately 61% of cervical cancer cases (consisting of a primary tumor and a germline
48 control) submitted to the BCR and processed passed quality control metrics. Tumor tissue from
49 173 cases was submitted for reverse phase protein array (RPPA) analysis.

50

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

62

63

64

65 **Sample Processing**

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

71

72 RNA samples were quantified by measuring Abs₂₆₀ with a UV spectrophotometer and DNA was
73 quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis
74 to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR
75 Identifiler (Applied Biosystems) was utilized to verify that tumor DNA and germline DNA
76 representing a case were derived from the same patient. Five hundred nanograms of each tumor
77 and normal DNA were sent to Qiagen (Hilden, Germany) for REPLI-g whole genome
78 amplification using a 100 µg reaction scale. RNA was analyzed via the RNA6000 nano assay
79 (Agilent) for determination of an RNA Integrity Number (RIN), and only analytes with RIN ≥7.0
80 were included in this study. Only cases yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg
81 RNA, and 4.9 µg of germline DNA were included in this study.

82

83 Samples with residual tumor tissue were considered for proteomics analysis. When available, a
84 10-20 mg piece of snap-frozen tumor adjacent to the piece used for molecular sequencing and
85 characterization was submitted to MD Anderson Cancer Center for RPPA analysis.

86 **Data Freeze**

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

90

91 **Histology Verification**

92 Frozen sections of all cervical cancers submitted for TCGA analysis were reviewed by a tissue site
93 pathologist and an independent pathologist prior to acceptance into the study. When available,
94 scanned images of the formalin-fixed, paraffin embedded tissue slides were reviewed by an expert
95 pathology panel. Only cases that met criteria for primary cervical cancer according to WHO
96 criteria⁶⁸ were accepted. These included squamous cell carcinomas, both large cell keratinizing in
97 which at least one well-formed keratin pearl was identified and large cell non-keratinizing.
98 Adenocarcinomas included adenocarcinoma of usual type, including mucin depleted, mucinous,
99 and endometrioid type. For analysis purposes, all adenocarcinomas were combined into one
100 endocervical adenocarcinoma category. Three adenosquamous carcinomas were also included.
101 All cervical cancers were assigned a pathologic grade, including Grade I: well-differentiated;
102 Grade II: moderately differentiated; and Grade III: poorly differentiated. Care was taken to verify
103 that the tumors included were not endometrial in origin.

104

105

106

107 **S2. HPV Detection and Integration**

108

109 **HPV Detection by MassArray (Nationwide Children’s Hospital)**

110 HPV status was determined by an ultra-sensitive method using real-time competitive polymerase
111 chain reaction and matrix-assisted laser desorption/ionization-time of flight mass spectroscopy
112 with separation of products on a matrix-loaded silicon chip array, similar to the work described in
113 Tang *et al*⁴⁵. Multiplex PCR amplification of the E6 region of 16 discrete high-risk HPV types
114 (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 90), 2 low-risk HPV types (HPV
115 6 and 11), and human GAPDH control was run to saturation followed by shrimp alkaline
116 phosphatase quenching. Amplification reactions included a competitor oligo identical to each
117 natural amplicon except for a single nucleotide difference. Probes that identify unique sequences
118 in the oncogenic E6 region of each type were used in multiplex single base extension reactions
119 extending at the single base difference between wild-type and competitor HPV so that each HPV
120 type and its competitor were distinguished by mass when analyzed on the MALDI-TOF mass
121 spectrometer.

122

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

124 The microbial detection pipeline used by the BC Cancer Agency’s Genome Sciences Centre (BC)
125 is based on BioBloom Tools (BBT, v1.2.4b1), which is a Bloom filter-based method for rapidly
126 classifying RNA-seq or DNA-seq read sequences⁴⁶. We generated 43 filters from “complete”
127 NCBI genome reference sequences of bacteria, viruses, fungi, and protozoa, using 25-bp k-mers

128 and a false positive rate of 0.02. We ran BBT in paired-end (PE) mode with a sliding window to
129 screen FASTQ files from RNA-seq libraries (48-bp PE reads, 178 tumors and no adjacent tissue
130 normals), and 40 whole genome shotgun libraries (WGS, 50-bp PE reads, 19 tumors and 19 blood
131 normals). In a single-pass scan for each library, BBT categorized each read pair as matching the
132 human filter, matching a unique microbial filter, matching more than one filter (multi-match), or
133 matching neither human nor microbe (no-match). For each filter, we then calculated a reads-per-
134 million (RPM) abundance metric as:

135

$$136 \text{ Abundance metric} = \left(\frac{\# \text{reads mapped to a microbe filter}}{\# \text{chastity passed reads in the sample}} * 10^6 \right)$$

137

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

145

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

147 The PathSeq algorithm⁴⁷ was used to perform computational subtraction of human reads, followed
148 by alignment of residual reads to a combined database of human reference genomes and microbial

149 reference genomes (which includes but is not limited to HPV genomes), resulting in the
150 identification of reads mapping to HPV genomes in RNA sequencing data.

151

152 Subjects were classified as HPV-positive by RNA sequencing if at least 1 HPV read in 1 million
153 human reads were present; otherwise, subjects were classified as HPV-negative. Using PathSeq,
154 human reads were subtracted by first mapping reads to a database of human genomes using BWA
155 (version 0.6.1)⁶⁹, Megablast (version 2.2.23), and Blastn (version 2.2.23)⁷⁰. Only sequences with
156 perfect or near perfect matches to the human genome were removed in the subtraction process. To
157 identify HPV reads, the resultant non-human reads were aligned with Megablast to a database of
158 microbial genomes that includes multiple HPV reference genomes. HPV reference genomes were
159 obtained from the NCBI nucleotide database (downloaded in June 2013).

160

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

162 An in-house developed pipeline, PathWatch, was used to detect bacteria and viruses and to
163 examine the integration status of the bacterial/viral genome. First, computational subtraction of
164 sequences mapped previously to the human genome was performed. Next, BWA was used to map
165 the remaining set of non-human sequences to the set of bacterial and viral reference genomes
166 obtained from the NCBI RefSeq database (<ftp://ftp.ncbi.nih.gov/refseq/release/microbial/> and
167 <ftp://ftp.ncbi.nih.gov/refseq/release/viral/> respectively). Reads that aligned to the genomes of
168 multiple species were filtered out. The percentage of covered pathogen genome, count of pathogen
169 sequencing reads normalized by the length of the pathogen genome, and total number of non-

170 human reads in the sample were calculated. To consider a given sample positive for the pathogen
171 presence we chose an empirical threshold of 1 kb of pathogen genome to be covered to distinguish
172 between positive calls and background noise from the reads that came from other species.

173

174 **HPV Variant Calling**

175 RNA-seq data in FASTA format was used to identify HPV variants (Supplemental Fig. S1).
176 Unaligned reads were taken from the PathSeq analysis (which contains HPV reads) and aligned to
177 HPV reference genomes (HPV complete genomes from NCBI) using TopHat⁴⁸ with default
178 parameters⁴⁹. A BAM file containing only the HPV-related reads was generated for each sample.
179 For each HPV isolate, a contig was generated using samtools⁷¹ and then aligned with the HPV
180 variant complete genome database⁷² to create a phylogenetic tree using RAxML⁷³. Single
181 Nucleotide Polymorphisms (SNPs) were called from the BAM file using samtools and SNVMix⁷⁴.
182 The HPV variant lineages/sublineages were assigned based on the phylogenetic topology by an in-
183 house script and confirmed visually using the SNP patterns⁵⁰.

184

185 **E6 Splicing Analysis**

186 The HPV splice junctions from RNA-seq were determined using TopHat. The splicing sites,
187 unspliced transcripts, and their prevalence were summarized with an in-house R script that
188 evaluated the RNA-seq reads within a window surrounding the splice sites within E6. Two
189 transcript types were distinguished for HPV16 and HPV18: (a) transcripts that included evidence
190 of an unspliced sequence of E6, and (b) a transcript spliced at the E6 splice donor site (position

191 226 for HPV16 and position and position 233 for HPV18) (spliced) (Supplemental Fig. S2). The
192 read counts for unspliced, spliced, and the sum of both transcript types, as well as the ratio of
193 unspliced/spliced transcripts were categorized into quartiles separately for HPV16 and HPV18
194 (Supplemental Table 3).

195

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

197 In order to assess potential genomic integration of HPV in 178 RNA-seq tumor libraries, ABySS
198 v1.3.4⁷⁵ was used to generate *de novo* assemblies for each library, using only the reads classified
199 by BBT (above) as human, HPV, multi-match, or no-match (Supplemental Fig. S3a). In order to
200 address how variations in transcript abundance influence assembly⁷⁶, we generated sets of
201 assemblies using every second k-mer length between 24 and 48 bp, and then generated a working
202 contig set for each library by merging the contigs from all of its k-mer assemblies using Trans-
203 ABySS v1.4.8⁷⁶. We reran BBT on the working contig set, applying only human and HPV filters
204 and identifying contigs that matched both filters. We identified viral-host chimeric contigs that
205 suggested splicing of HPV donor splice sites into host splice acceptor sequences by using BLAT
206 v34⁷⁷ to align each contig to the GRCh37-lite human reference genome and to 293 HPV reference
207 genomes. After removing any human/viral contig that had a gap longer than 10 bp between the
208 human- and viral-aligned segments, we retained the highest-scoring human-viral contig alignment
209 combination. We required a contig's aligned sequences to span at least 90% of its overall length,
210 and to overlap by less than 50%. We required a viral-human contig junction to have at least 5
211 mate flanking reads or 3 mate spanning reads (Supplemental Fig. S4a, b). Human splice junction

212 contig coordinates were annotated against RefSeq and UCSC gene annotations (last modified on
213 June 30, 2013) from the UCSC genome browser⁷⁸.

214

215 Since the chimeric contig junctions represent splicing between a viral transcript and a human
216 transcript, the junction coordinate in each genome may not correspond to the actual location of the
217 DNA integration, and a given genomic (i.e. DNA) integration event can be reported in RNA-seq
218 data as multiple transcript splice sites whose genomic locations span large distances⁷⁹.

219

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

221 An HPV-positive sample was considered integration positive if there were at least 5 flanking reads
222 and 10 total spanning reads (summing mate and single) supporting an integration site. Flanking
223 read pairs were defined as having one end of the paired-end read mapped to the HPV genome and
224 its mate pair mapped to the human genome. Spanning reads were defined as having one end of
225 the paired-end read spanning the integration junction and its mate pair mapped to either the human
226 or HPV genome. Once HPV reads were obtained (Pathseq, above), we extracted all pair mates
227 and used Tophat-2.0.8⁸⁰ with fusion option enabled to map these paired end reads to a combined
228 database containing the human genome and an HPV genome. Next, spanning reads and flanking
229 reads were identified from the aligned BAM file. Human genes involved in the integration were
230 identified using the breakpoint coordinates against RefSeq and UCSC gene annotations (last
231 modified on June 30, 2013) from the UCSC genome browser⁷⁸.

232 **Inter-Center Concordance Calls for RNA-seq Integration Events**

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

237

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

244

245 Consistent with sets of chimeric viral-human transcripts being derived from a genomic integration
246 location, we noted that sites reported by both methods tended to occur as localized clusters. Given
247 this clustering, sites on each chromosome were combined into a smaller set of 'events' using a
248 500-kb window and locating an event at the midpoint of its supporting sites (Supplemental Fig.
249 S4c). The events identified by assembly were then compared with those from read alignments on
250 both the patient and event levels (Supplemental Fig. S5).

251

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

261

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

267

268 For the 169 HPV-positive patients, 141 patients had integration events that were confident or
269 inferred-confident, while the remaining 28 patients had no confident integration events. Of the
270 141 patients, 129 had events called by both methods, two had confident events that were called
271 only by BC, and 10 had confident events called only by BI. Of the 129 patients with events called
272 by both methods, all events were concordant in 90 patients. These concordant events consisted of

273 91 that were confident and 6 that were inferred confident. In 39 of the 129 patients, there were
274 both concordant and discordant events. These events consisted of 43 concordant/confident events,
275 4 concordant/inferred confident, 1 concordant and not confident, and 57 (12 BC and 45 BI) that
276 were discordant and confident.

277

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

279 A pipeline was used that took advantage of paired-end (PE) sequencing technology and searched
280 for the clusters of discordant read pairs where one mate is aligned to the human genome and the
281 second mate mapped to the viral sequence. As an input, an original set of all PE reads that was
282 mapped and unmapped to the human genome was used. Two subsets of reads were generated:
283 ends represented by human sequences and their unmapped mates. Such unmapped reads were then
284 aligned against the specific viral genome identified in the previous step. Clusters of discordant
285 read pairs were calculated. In order to determine the presence of a cluster, we used an empirical
286 cutoff of 3 discordant read pairs within the same integration region. Chimeric viral-human reads
287 were then searched to assess the precise site of a candidate integration event at nucleotide
288 resolution. Soft-clipped reads, in which only a portion of a read had been mapped to the human
289 genome, were filtered from the original PE dataset and were aligned by BLAT (v.34) to the virus
290 genome.

291

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

293 WGS data for 70 tumor samples were downloaded from CGHub and aligned to a custom reference
294 consisting of human GRCh37-lite and HPV 6, 16, 18, 31, 33, 35, 39, 45, 52, 56, 58, and 59
295 sequences, along with Polyoma BK, Polyoma HPyV7, Hepatitis B, Merkel Cell Polyoma as well
296 as HHV 1, 4, and 5. Bwa v0.5.9 was used with default parameters for both bwa aln and bwa
297 samse/sampe, using bwa's built in quality-based read trimming (-q 5).

298
299 Virus and discordant reads were discovered by parsing the realigned BAM using samtools (version
300 0.1.18) and standard UNIX utilities. Virus reads were detected in 66 samples, and discordant reads
301 were observed in 65 samples. Sixty-three samples with 5 or more discordant reads were analyzed
302 with Pindel version 0.2.5a2⁸¹ read pair (RP) module, and human-virus breakpoints were observed
303 for 44 of these. Breakpoint position is returned as a range of positions on both human chromosome
304 and virus, with accuracy limited by insert size to approximately ± 1000 bp.

305 306 **Integration Analysis with Copy Number, mRNA Expression, and Structural Variant Data**

307 We assessed gene-level expression relative to somatic copy number and structural variant data for
308 genes into which we had mapped viral-human junctions from RNA or DNA sequencing data, and
309 for genes that were associated with enhancers into which we had mapped RNA or DNA junctions.
310 We used somatic copy number from a GISTIC2.0 "all_data_by_genes.txt" file, and normalized
311 RSEM gene-level RNA-seq data. We assessed viral strain, viral splice donor and acceptor
312 coordinates⁸², and total read evidence for viral-human splice junctions, considering read evidence

313 separately for the two methods. From the combined RNA and DNA evidence, we generated
314 schematic splicing diagrams involving viral and human transcripts.

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

321

322

323

324

325

326

327

328

329

330

331

332

333

334 **S3. DNA Sequencing and Mutation Calling**

335

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

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

345

346 **Read Duplication Marking and Merging**

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

355 **Low-Pass WGS Sequencing Methods**

356 Between 500 and 700 ng of each gDNA sample were sheared to approximately 250 bp fragments
357 using Covaris E220 and then converted to a pair-end Illumina library using KAPA Bio kits with
358 Caliper (PerkinElmer) robotic NGS Suite according to manufacturers' protocols. All libraries
359 were sequenced on HiSeq2000 using one sample per lane, with the pair-end 2 x 51bp setup. Tumor
360 and its matching normal were usually loaded on the same flow cell. Raw data were converted to
361 the FASTQ format and BWA alignment was used to generate bam files.

362

363 **Somatic Mutation Calling**

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

368

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

375

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

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

379

380 **BCGSC:** `bcgsc.ca_CESC.IlluminaHiSeq_DNASeq_automated.Level_2.1.0.0.tar.gz`

381 **UCSC:** `ucsc.edu_CESC.IlluminaGA_DNASeq_automated.Level_2.1.0.0.tar.gz` (Single

382 nucleotide somatic mutations were identified by RADIA (RNA and DNA Integrated Analysis))⁸³

383 **Broad:** `broad.mit.edu_CESC.IlluminaGA_DNASeq_automated.Level_2.1.6.0.tar.gz`

384

385 Readcounts supporting the tumor and variant allele for all predicted somatic mutations were
386 extracted from exome BAM pairs using `bam-readcount v0.5` (<https://github.com/genome/bam-readcount>).
387

388

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

394

395 Potential false positives due to germline cross contamination were removed by filtering all
396 germline variants from dbSNP 137 VCF files with a $GMAF > 0$. In order to obtain a set of high

397 confidence somatic variants, the following minimum supporting requirements were set: Minimum
398 tumor supporting reads ≥ 2 , minimum tumor VAF of 10%, minimum normal reference supporting
399 reads ≥ 8 , and maximum normal variant supporting reads ≤ 1 .

400
401 Previously identified, recurrent false positives identified in other TCGA exome data were filtered
402 as previously described⁸⁴ and remaining novel recurrent somatic mutations were manually curated
403 to identify and remove further artifacts.

404

405 **Identifying Significantly Mutated Genes (SMGs)**

406 Mutations were included for the Extended set of 192 samples, with 178 being part of the Core
407 Freeze set. Eleven samples were identified to exhibit greater than average mutations rates and
408 were termed “hypermutants” (somatic mutations >600). These 11 samples were removed when
409 identifying SMGs. MutSig⁶ was utilized to identify SMGs within the exome sequencing data. All
410 3 sample subsets without “hypermutants” (Supplemental Table 4) were analyzed using an FDR
411 cutoff of 0.1. Significant p-values and FDR values are shown in Supplemental Table 4.

412

413 **Somatic Mutation and Structural Variant Validation Methods**

414

415 **Library Hybrid Capture**

416 Tumor and normal Illumina libraries were enriched by performing hybrid capture using Roche
417 Nimblegen SeqCap EZ custom capture oligos. Genomic DNA was utilized for library construction

418 starting material when available, and Qiagen WGA amplified DNA was used when insufficient
419 material was available. Each sample library received unique, dual molecular barcodes prior to
420 pooling. The target regions for somatic indels and point mutations were the 100bp region
421 surrounding the mutation site, while for RNA-seq fusion transcript validation the flanking region
422 of the largest introns flanking each novel exon-exon junction were targeted. Probes designed with
423 >5 mismatches were discarded. Additional 120-mer IDT probes targeting cancer-related viruses
424 were combined with SeqCap custom probes prior to capture. Target and probe bed files are
425 available at http://genome.wustl.edu/pub/custom_capture/. Each sample was pooled into one of
426 ten sets, each containing 40 or 41 samples. Each set was captured independently and sequenced
427 on one lane of Illumina HiSeq 1T with an estimated target coverage of 200-300x.

428

429 **Read Alignment**

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

437

438

439 **Read Duplication Marking and Merging**

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

448

449 **Somatic Variant Calling**

450

451 **SNV Callers**

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

456

457

458

459 **SNV Caller Combination and Filtering**

460 First, Samtools calls were retained if they met all of the following rules inspired by MAQ: Site is
461 greater than 10 bp from a predicted indel of quality 50 or greater, the maximum mapping quality
462 at the site is ≥ 40 , fewer than 3 SNV calls in a 10 bp window were around the site, site is covered
463 by at least 3 reads and less than 1,000,000,000 reads, consensus quality ≥ 20 , and SNP quality \geq
464 20.

465
466 After these filters were applied, Samtools and SomaticSniper calls were unioned using joinx v1.9
467 (<https://github.com/genome/joinx>; `joinx sort --stable --unique`). The resulting merged set of
468 variants were additionally filtered to remove likely false positives. Bam-readcount v0.4
469 (<https://github.com/genome/bam-readcount>) was used with a minimum base quality of 15 (-b 15)
470 to generate metrics and retained sites based on the following requirements: Minimum variant base
471 frequency at the site of 5%, percent of reads supporting the variant on the plus strand $\geq 1\%$ and \leq
472 99% (variants failing these criteria were filtered only if the reads supporting the reference did not
473 show a similar bias), minimum variant base count of 4, variant falls within the middle 90% of the
474 aligned portion of the read, maximum difference between the quality sum of mismatching bases
475 in reads supporting the variant and reads supporting the reference of 50, maximum mapping quality
476 difference between reads supporting the variant and reads supporting the reference of 30,
477 maximum difference in aligned read length between reads supporting the variant base and reads
478 supporting the reference base of 25, minimum average distance to the effective 3' ends of the read

479 for variant supporting reads of 20% of the sequenced read length, and maximum length of a
480 flanking homopolymer run of the variant base of 5.

481

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

485

486 VarScan calls were retained if VarScan reported a somatic p-value ≤ 0.07 , a normal frequency \leq
487 5%, a tumor frequency $\geq 10\%$, and ≥ 2 reads supporting the variant. VarScan variants passing
488 these criteria were then filtered for likely false positives using bam-readcount v0.4 and identical
489 criteria as described above for SomaticSniper. Fully filtered calls as described above for
490 SomaticSniper and VarScan were then merged with calls from Strelka using joinx v1.9 (joinx sort
491 --stable --unique) to generate the final callset.

492

493 **Indel Callers**

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

499

500 **Indel Caller Combination and Filtering**

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

511

512

513

514

515

516

517

518

519

520

521 **S4. Copy Number Variation (CNV) Analysis**

522

523 **CNV Methods**

524 DNA processing via SNP 6.0 arrays is described in Methods. Briefly, Birdseed was used to infer
525 a preliminary copy number at each probe locus from raw .CEL files⁵². For each tumor, genome-
526 wide copy number estimates were refined using tangent normalization, in which tumor signal
527 intensities are divided by signal intensities from the linear combination of all normal samples that
528 are most similar to the tumor¹⁶. This linear combination of normal samples tends to match the
529 noise profile of the tumor better than any set of individual normal samples, thereby reducing the
530 contribution of noise to the final copy number profile. Individual copy number estimates then
531 underwent segmentation using Circular Binary Segmentation⁵³. As part of this process of copy
532 number assessment and segmentation, regions corresponding to germline copy number alterations
533 were removed by applying filters generated from TCGA germline samples from the ovarian cancer
534 analysis and from samples of this cohort. Segmented copy number profiles for tumor and matched
535 control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously
536 assigns a length and amplitude to the set of inferred copy number changes underlying each
537 segmented copy number profile⁵⁴. Significance of copy number alterations were assessed from
538 the segmented data using GISTIC2.0 (Version 2.0.22)⁵⁴. Briefly, GISTIC2.0 deconstructs somatic
539 copy number alterations into broad and focal events and applies a probabilistic framework to
540 identify location and significance levels of somatic copy number alterations.

541

542 **Results**

543 Somatic copy number alterations in 178 CESC tumors were determined with SNP 6.0 arrays. There
544 were an average of 88 copy number alterations per tumor, less than ovarian and serous endometrial
545 carcinomas but more than endometrioid endometrial carcinomas^{16,17}. Analysis of focal
546 amplifications and deletions performed by the GISTIC2.0 algorithm revealed 26 focal
547 amplifications and 37 focal deletions along with 23 whole arms that were recurrently altered.
548 Recurrent focal amplifications were identified at 3q26.31 (*TERC*, *MECOM*), 3q28 (*TP63*), 7p11.2
549 (*EGFR*), 8q24.21 (*MYC*, *PVT1*), 9p24.1 (*CD274*, *PDCDILG2*), 11q22.1 (*YAP1*), 13q22.1 (*KLF5*),
550 16p13.13 (*BCAR4*), and 17q12 (*ERBB2*). Recurrent deletions were identified at 4q35.2 (*FAT1*),
551 3p24.1 (*TGFBR2*), 10q23.31 (*PTEN*), and 18q21.2 (*SMAD4*). Notably, this analysis discovered
552 novel cervical cancer driver genes, including the therapeutic targets of immune inhibitors *CD274*
553 (*PD-L1*), *PDCDILG2* (*PD-L2*), and novel linc-RNA *BCAR4*. The amplifications of *PDL1/2*
554 correlated significantly ($p < 0.0001$) with cytolytic activity¹⁸. *BCAR4*, which has been
555 characterized for its role in promoting metastasis, anti-estrogen resistance, and Lapatinib
556 sensitivity in breast cancer¹⁹, was highly amplified, fused, and greatly overexpressed compared to
557 other tumors that do not express the gene.

558 Unsupervised clustering of somatic copy number alterations revealed two groups of
559 tumors, one group with a high rate of copy number alterations and one with less ($p < 0.0001$).
560 Interestingly, these groups also showed significant clinical and molecular differences. The CN
561 high cluster was largely composed of squamous tumors infected with HPV16 and contained
562 significantly more tumors with *YAP1* amplifications ($p < 0.0001$). The CN low cluster contained

563 the majority of adenocarcinomas, HPV18-infected samples, and presented a novel deletion of
564 *TGFBR2* as well as gains of *BCAR4* and *PDL1/2*.

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585 **S5. mRNA Sequencing, Analysis, and Structural Variants**

586

587 **RNA-seq Methods**

588 RNA was processed as described in Methods. For further details on this processing, refer to
589 Description file at the DCC data portal under the V2_MapSpliceRSEM workflow ([https://tcga-](https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illumina)
590 [data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illumina](https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illumina)
591 [hiseq_rnaseqv2/rnaseqv2/unc.edu_CESC.IlluminaHiSeq_RNASeqV2.mage-](https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illumina)
592 [tab.1.9.0/DESCRIPTION.txt](https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/cesc/cgcc/unc.edu/illumina)).

593

594 **Unsupervised Expression Clustering**

595 Genes with >10% missing normalized RSEM values across samples were removed from the Core
596 Freeze dataset (n=178 samples). RSEM values were then log₂-transformed after first adding a
597 constant of 1 to all values. The gene expression matrix was further filtered to only include the top
598 10% most variable genes by mean absolute deviation (n=1176 genes). Consensus clustering using
599 self-organized maps was employed to identify the most robust expression clusters for between 2
600 to 6 clusters. Rank survey profiles for the cophenetic and silhouette widths, along with consensus
601 cluster membership heatmaps (data not shown) suggested that a 3-cluster solution was optimal. A
602 nearest centroid-based classifier (CLaNC) was used to identify a set of signature genes which had
603 the lowest cross validation and prediction errors for sample membership in their respective
604 clusters⁸⁵. Hierarchical clustering was performed after median-centering gene expression values

605 using Cluster 3.0⁸⁶ (uncentered correlation with centroid linkage) and visualized using
606 JavaTreeview⁸⁷.

607

608 **Identifying a Uterine Corpus Endometrial Carcinoma (UCEC) Gene Classifier**

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

623

624

625

626 **Comparing CESC, UCEC, and HNSC Gene Expression Profiles**

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

636

637 **Detecting Structural Variants from RNA-seq and WGS Data**

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

641

642 *A. TopHat-Fusion and BreakFusion*

643 We ran Tophat-fusion-0.1.0 (Beta)⁸⁸ and BreakFusion-1.0.1⁸⁹ on each of the BAM files for the
644 Core Set samples to identify fusion candidates. We further filtered the identified candidates if a)
645 the gene fusion pairs were identified in the normal RNA libraries in the 1000 Genomes project⁹⁰;

646 b) the fusion breakpoints were 10 bp or more away from known splicing sites in the Refseq
647 database; or c) they were in self-chain regions with a self-chain alignment score greater than 10.

648

649 B. *PRADA*

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

665

666 C. *MapSplice*

667 RNA-seq data was processed and analyzed using MapSplice version 2.0.1.9 for potential gene
668 fusions as previously described¹⁰ to decrease the number of false positives. The resulting gene
669 fusion list was manually curated and filtered to only include potential events where both the donor
670 and acceptor sequences lie within known genes. To increase the confidence in the called fusions,
671 the list of potential gene fusions was further refined to include only fusions with coverage of at
672 least 10 reads and that had at least 2 reads bridging the breakpoint.

673

674 Detection of structural variations in low-pass WGS data (n=50) was performed using two
675 algorithms: BreakDancer⁵⁶ and Meerkat⁵⁷. The first step in BreakDancer requires a configuration
676 file of each BAM file for each tumor pair with the bam2cfg.pl perl module of the program. The
677 perl module BreakDancerMax.pl is then run on the configuration file to call structural variants in
678 the tumor and control files. The set of structural variant calls from each tumor sample was filtered
679 by the calls from its matched normal to remove germline variants. Structural variations were also
680 detected by Meerkat, which requires at least two discordant read pairs supporting each event and
681 at least one read covering the breakpoint junction. Variants detected from tumor genomes were
682 filtered by the variants from all normal genomes to remove germline events and were also filtered
683 out if both breakpoints fell into simple repeats or satellite repeats. The final call needed to fulfill
684 the following: (1) the read identified to span the breakpoint junction hit the predicted breakpoint
685 region uniquely by BLAT; or (2) the mate of the read spanning the breakpoint junction was mapped
686 near the predicted breakpoint.

687

688 High-pass WGS data (n=19) were analyzed to detect somatic structural variations using two runs
689 of BreakDancer and one run of SquareDancer (<https://github.com/ding-lab/squaredancer>). The
690 predictions were unioned after filtering each set of predictions with TigraSV⁹³, assembly-based,
691 and breakpoint confirmed. To detect interchromosomal breakpoints, Breakdancer 1.4.2 was run
692 with the optional parameters "-g -h:-a -t -q 10 -d". To detect intrachromosomal breakpoints,
693 Breakdancer 1.4.2 was run with the optional parameters "-g -h:-a -q 10 -o". Squaredancer v0.1
694 was run with default parameters.

695
696 Gene fusion lists generated by all methods and platforms were integrated. We identified 22
697 putative structural rearrangements detected by both RNA-seq and WGS (Supplemental Table 8).
698 In total, 26 recurrent fusions were identified, of which 3 were detected by at least two RNA-seq
699 callers (Supplemental Table 9). Furthermore, for the samples that did not have WGS data
700 available, we extended the analysis performed on the PRADA RNA-seq fusion calls on SNP6
701 array copy number data to any junction points predicted by all three RNA callers described above.
702 Among those, 74 fusions were detected by at least 2 RNA-seq callers and 60 of them showed
703 supporting breakpoints existing within 100 kb in SNP6 array data (Supplemental Table 10).

704

705 **mRNA Results**

706 Consensus clustering was performed on RNA-seq data from 178 CESC tumor samples using 1,176
707 highly variable genes to identify stable subgroupings of samples. Based on this expression data,
708 the cervical cancer samples were separated into 3 stable clusters. A gene signature was developed

709 consisting of 300 genes which performed optimally for grouping the samples into the clusters
710 identified by consensus clustering. Hierarchical clustering using centroid linkage resulted in the
711 samples being grouped into 3 clusters (Supplemental Fig. S9). Functional gene annotation analysis
712 and gene set enrichment analysis were used to identify the biologic processes involved in the
713 separation of the cervical cancer samples into the 3 clusters. Samples in Cluster C1 contained all
714 but 1 of the adenocarcinomas and 2 of 3 adenosquamous samples, suggesting that this is the Non-
715 Squamous cluster. Interestingly, this cluster also includes 15 squamous cell carcinomas with
716 expression patterns more closely related to the non-squamous cell cancers. Samples in this cluster
717 exhibit increased expression in genes such as *EPCAM*, *CLDN3*, *ERBB4*, *RAB17*, and *KRT18*, while
718 also showing markedly reduced expression of genes encoding several small proline-rich proteins
719 (SPRRs), p63, and FAT2. Samples in Cluster C2 consisted entirely of squamous cell carcinomas.
720 Genes with elevated expression in this cluster showed enrichment of ectoderm development genes
721 and cell junction genes. Representative genes with elevated expression include 8 members of the
722 keratin family, *ZNF750*, and *APOBEC3A*. The robust expression of keratin family member genes
723 suggests that this cluster could be considered a Squamous Cell – Keratinizing cluster. Samples in
724 Cluster C3 consisted entirely of squamous cell carcinomas, with the addition of 1 adenocarcinoma
725 and 1 adenosquamous sample. Genes with elevated expression in this cluster showed enrichment
726 of glycoprotein genes such as *EPHB2* and *TGFB2*. Samples in this cluster generally have lower
727 expression of keratin family members, suggesting that this cluster could be considered the
728 Squamous Cell – Non-Keratinizing cluster.

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

746

747 ***Cervical cancer/Endometrial cancer classification:*** Since primary cervical cancers can be
748 confused with endometrial cancers that involve the cervix secondarily, we developed a gene

749 expression classifier that differentiated cervical cancers from endometrial cancers. After randomly
750 sorting the cervical and endometrial cancer samples into a training and test set, a 14 gene classifier
751 was identified that had the lowest prediction error in the training set, with 0 (0%) classification
752 errors for the endometrial samples and 4 (4.4%) classification errors for the cervical samples, for
753 an overall error rate of 2.3%. Similar results were observed when applied to the test set: 0 (0%)
754 classification errors for the endometrial samples and 4 (4.3%) classification errors for the cervical
755 samples, for an overall error rate of 2.3%. These 8 cervical cancer samples predicted to be
756 endometrial cancers by expression profiling were reevaluated by study pathologists who confirmed
757 that these samples did indeed arise from the cervix, thus we term these samples as endometrial-
758 like (UCEC-like) cervical cancers. Interestingly, these 8 endometrial-like cervical cancers include
759 7 of the 9 HPV negative cancers and all but 2 of the cancers have a non-squamous cell histology.
760 Next, gene expression profiles were compared between cervical cancers and endometrial cancers
761 by identifying differentially expressed genes between the 2 cancer types. Unsurprisingly, the
762 cervical and endometrial cancers tended to cluster among members of the same tissue type, except
763 for 6 of the 8 endometrial-like cervical cancers, which clustered among the endometrial cancers.
764 The other 2 endometrial-like cervical cancers clustered with the C1 cervical cancers, along with 1
765 endometrial cancer sample (Supplemental Fig. S10).

766

767 ***Cervical/Endometrial/Head and Neck cancer comparison:*** Gene expression profiles were
768 compared between cervical (CESC), endometrial (UCEC), and head and neck cancers (HNSC).
769 Hierarchical clustering of the different cancer samples across 4,039 highly variable genes

770 separated the samples predominantly according to cancer type, with a few exceptions. The cervical
771 adenocarcinomas tend to congregate in a subcluster, along with the other samples in the non-
772 squamous expression clusters samples, and have expression patterns quite similar to those of
773 UCEC samples. A group of about 700 genes with relatively greater expression are shared between
774 the CESC samples in this subcluster and UCEC samples in general. Functional analysis of these
775 genes shows overrepresentation of genes involved in embryonic morphogenesis (*HOXA9*,
776 *HOXB2-9*) and the axoneme (6 members of the dynein family). In addition, this group of samples
777 exhibit elevated expression of genes seen in the Non-Squamous CESC expression cluster (*ERBB4*,
778 *RAB17*, *KRT18*) and genes highly expressed in UCECs (*ESR1* and *PGR1*). Further, a group 27
779 HNSC samples grouped within the CESC cluster. Interestingly, 23 of these samples are HPV-
780 positive compared to only 13 out of 256 samples in the HNSC cluster ($p < 0.0001$; Fisher's Exact
781 test). Functional analysis of the gene expression patterns shared by HPV-positive HNSC and
782 CESC samples may provide insights into the effects of HPV in oncogenesis. The analysis of
783 shared genes with relatively increased expression resulted in an overrepresentation of genes
784 involved in meiosis, including *MEI1*, *STAG3*, *SYCEP2*, and *SYCP2* which have previously been
785 shown to be increased in HPV-positive cancers. In addition, the HPV-positive HNSC samples that
786 group in the CESC cluster show decreased expression of a large number of genes that exhibit
787 increased expression in the HNSC cluster. Functional analysis of these genes show
788 overrepresentation of genes involved in ectoderm development, cell adhesion, serine-protease
789 inhibitor activity, wound healing, and angiogenesis (Extended Data Fig. 4a).

790

791 **Structural Variant Results**

792 To characterize structural rearrangements we performed an integrative analysis of RNA-seq
793 (n=178) and WGS data with low-pass (n=50) and deep (n=19) coverage. We identified 22 putative
794 structural rearrangements detected by both RNA-seq and WGS (Supplemental Table 8). In total,
795 26 recurrent fusions were identified, of which 3 were detected by at least two RNA-seq callers
796 (Supplemental Table 9). Examples of putative driver events are a *FGFR3-TACC3* fusion (n=1),
797 already known in other cancer types^{9,94} but not previously reported in cervical cancer, and
798 *ZC3H7A-BCAR4* fusions (n=4). These fusions linked exon 1 of *ZC3H7A* to exon 4 of *BCAR4*. The
799 long non-coding RNA *BCAR4* has been shown to promote estrogen-independent growth and
800 tamoxifen resistance in breast cancer^{95,96}.

801

802

803

804

805

806

807

808

809

810 **S6. Methylation Analysis**

811

812 **Sample Preparation and Hybridization**

813 The Illumina Infinium HM450 array⁵⁸ was used to assay the Core Set of 178 TCGA cervical cancer
814 samples. This platform includes probes for more than 480,000 CpG sites, spanning 99% of RefSeq
815 genes. In total, 96% of CpG islands and 92% of CpG shores are represented by at least one probe.
816 Genomic DNA (1000 ng) for each sample was treated with sodium bisulfite, recovered using the
817 Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA) according to the manufacturer's
818 specifications, and eluted in an 18 μ L volume. All TCGA DNA samples passed quality control
819 and proceeded to the Infinium DNA methylation assay. Each bisulfite-converted DNA sample
820 was whole genome amplified (WGA) followed by enzymatic fragmentation as specified by the
821 manufacturer. The bisulfite-converted, fragmented WGA-DNA samples were then hybridized
822 overnight to 12 sample BeadChips. During this hybridization, the WGA-DNA molecules anneal
823 to methylation-specific DNA oligomers linked to individual bead types, with each bead type
824 corresponding to a specific DNA CpG site and methylation state. The oligomer probe designs
825 follow the Infinium I and II chemistries, in which locus-specific base extension follows
826 hybridization to a methylation-specific oligomer. There are two different bead types for each
827 locus, one with an oligomer that anneals specifically to the methylated version of the locus and the
828 other with an oligomer that anneals to the unmethylated version of the locus. The Infinium I probes
829 terminate complementary to the interrogated CpG site for methylated loci, or complementary to
830 the TpG for unmethylated alleles. A matched oligomer-template DNA molecule hybrid will allow

831 for the incorporation of a labeled nucleotide immediately adjacent to the interrogated CpG (or
832 TpG) site. However, if the probe and template are mismatched, then primer extension will not
833 occur. Adenine and thymine nucleotides are labeled with cy5 (red), while cytosine nucleotides are
834 labeled with cy3 (green). No insertion of guanine nucleotides occurs in Infinium I assays. Of
835 note, the identity of the dye is representative of the nucleotide adjacent to the CpG dinucleotide.
836 The methylation discrimination is derived from separate measurements from the two different
837 types of beads present for each locus. For some loci, both measurements will be cy3, and for
838 others both will be cy5. The Infinium type II chemistry is a true two-color system. A matched
839 oligomer-template DNA molecule hybrid will allow for the incorporation of a labeled nucleotide
840 immediately 3' to the interrogated CpG (or TpG) site. Adenine nucleotides labeled with cy5 (red)
841 are incorporated at unmethylated (TpG) sites, while guanine nucleotides labeled with cy3 (green)
842 are incorporated at methylated (CpG) sites. The intensities of both cy3 and cy5 are obtained after
843 scanning each hybridized array. BeadArrays are scanned and the raw data are imported into
844 custom programs in R computing language for pre-processing and calculation of beta value DNA
845 methylation scores for each probe and sample.

846

847 **Data Processing**

848 Probes having a common single nucleotide polymorphism (SNP) (defined as a SNP with a minor
849 allele frequency > 1% as defined by the UCSC snp135common track) within 10 bp of the
850 interrogated CpG site and probes that overlapped with a REPEAT element (as defined by
851 RepeatMasker and Tandem Repeat Finder Masks based on UCSC hg19, Feb 2009) within 15 bp

852 of the interrogated CpG site were identified and excluded from subsequent analyses. In addition,
853 probes with a non-detection probability (detection p-value) greater than 0.05 in more than 25% of
854 the samples and those associated with the Y chromosome were excluded. Probes that are mapped
855 to multiple sites on hg19 were annotated as NA for chromosome and 0 for CpG/CpH coordinate.
856 The final number of probes after the above exclusions was 395,552 probes.

857

858 The Illumina HumanMethylation450 array uses two different types of probes. Specifically, this
859 array profiles the methylation status of 485,577 CpG dinucleotides, of which 72% use the Infinium
860 type II primer extension assay where the unmethylated (red channel) and methylated (green
861 channel) signals are measured by a single bead⁵⁸. The remainder use the Infinium type I primer
862 extension assay (also used in the Illumina HumanMethylation27 array) where the unmethylated
863 and methylated signals are measured by different beads in the same color channel. Importantly,
864 the two probe types differ in terms of CpG density, with more CpGs mapping to CpG islands for
865 type I probes (57%) as compared with type II probes (21%). Moreover, compared with Infinium
866 I probes, the range of beta values obtained from the Infinium II probes is smaller. In addition, the
867 Infinium II probes have also been shown to be less sensitive for the detection of extreme
868 methylation values and display a greater variance between replicates⁹⁷.

869

870 **Clustering Analysis**

871 Unsupervised consensus clustering was performed as implemented in the Bioconductor package
872 ConsensusClusterPlus, with Euclidean distance and partitioning around medoids (PAM).

873 Consensus clustering was applied to the DNA methylation data from the entire cohort, using the
874 most variable 1% of CpG island promoter probes.

875

876 **Identification of Epigenetically Silenced Genes**

877 Epigenetically silenced genes were identified as previously described⁵⁹. Specifically, we first
878 identified promoter CpG sites that met several criteria: (a) at least 90% of normal samples should
879 be clearly unmethylated ($\beta \leq 0.1$) at that site; (b) at least 5% of tumor samples should be clearly
880 methylated ($\beta \geq 0.3$) at that site; and (c) a t-test comparing expression levels in methylated ($\beta \geq 0.3$)
881 and unmethylated tumor samples ($\beta < 0.1$) should be significant at an $FDR < 0.01$. A gene was
882 defined as epigenetically silenced if at least 25% of the promoter CpG sites met all of these criteria.
883 A total of 120 normal samples were used for this analysis, including 10 each drawn at random
884 from the 12 TCGA projects that include normal samples, such as lung adenocarcinoma⁹⁸, breast
885 invasive carcinoma¹¹, colon adenocarcinoma⁹⁹, endometrial carcinoma¹⁷, and others. Fisher's
886 Exact test was used to find pathways enriched with epigenetically silenced genes. Pathways with
887 $FDR < 0.05$ were considered significantly enriched.

888

889 **HPV DNA Methylation Signatures**

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

894 as described¹⁰. Additionally, empirical Bayes moderated T-tests¹⁰⁰ were used to identify
895 methylation differences between HPV clades A7 and A9.

896

897 **Additional Analyses**

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

904

905 **Results**

906 Classifications with 2 to 7 groups were evaluated for cluster stability and fit to choose a final
907 partition of the samples. The DNA methylation based subtypes presented here are based on a
908 robust 3-group partition of the samples obtained using the most variable CpG island promoter
909 features on the Illumina Infinium HM450 array (Extended Data Fig. 5). A CIMP-high (CpG island
910 hypermethylated) cluster is characterized by widespread methylation at CpG sites within gene
911 promoter and CpG island regions, while the CIMP-low group is distinguished by very little
912 methylation within islands, a methylation pattern typical of healthy epithelial tissue. HPV- tumors
913 formed a distinct cluster within the CIMP-low group with a significantly lower mean promoter
914 methylation level than the rest of the samples in that group (t-test p-value = 0.005). The CIMP-

915 high cluster contained most of the endocervical adenocarcinoma samples and was enriched with
916 samples from mRNA cluster 1, miRNA cluster 4, CN-low cluster, and the Adenocarcinoma
917 iCluster. In addition, this cluster had higher purity samples with lower EMT score as shown by
918 boxplots in Extended Data Fig. 5b. There was no significant difference in survival between the
919 methylation clusters (log-rank test $p=0.9$)

920

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

927

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

934

935

936 **S7. microRNA Sequencing and Analysis**

937

938 **Libraries and Sequencing**

939 MicroRNA sequence (miRNA-seq) data was generated for the Core Set of 178 tumor samples
940 using methods described previously¹¹. Reads were aligned to the GRCh37/hg19 reference human
941 genome and read count abundance was annotated against miRBase v16 stemloops and mature
942 strands using only exact-match read alignments. Of note, BAM files that include all sequence
943 reads are available from CGHub (cghub.ucsc.edu)¹⁰¹. miRBase v20 was used to assign 5p and 3p
944 mature strand (miR) names to MIMAT accession IDs.

945

946 **Unsupervised Clustering**

947 Groups of samples that had similar abundance profiles were identified using unsupervised non-
948 negative matrix factorization (NMF) consensus clustering (v0.20.5) in R 3.1.2, with default
949 settings¹⁰². The input was a reads-per-million (RPM) data matrix for the 303 (25%) most-variant
950 5p or 3p mature strands. After running a rank survey with 50 iterations per solution, we chose a
951 preferred clustering solution and performed a 500-iteration run to generate the final clustering
952 result. The preferred solution was chosen by considering profiles of the cophenetic correlation
953 coefficient and the average silhouette width calculated from the consensus membership matrix,
954 Kaplan-Meier survival analysis, and clinical covariate associations for a range of candidate
955 clustering solutions. To visualize typical vs. atypical cluster members, a profile of silhouette

956 widths was calculated from the final NMF consensus membership matrix, whereby atypical cluster
957 members have relatively low widths.

958

959 To generate a heatmap for the NMF results, we first identified miRs that were differentially
960 abundant between the unsupervised miRNA clusters using a SAMseq multiclass analysis (samr
961 2.0)¹⁰³ in R with a read-count input matrix and an FDR threshold of 0.05. For the heatmap, miRs
962 that had the largest SAMseq scores and median abundances greater than 25 RPM were included.
963 The RPM filtering acknowledged potential sponge effects from competitive endogeneous RNAs
964 (ceRNAs) that can make weakly abundant miRs less influential^{104,105}. Each row of the matrix was
965 transformed by $\log_{10}(\text{RPM} + 1)$ and then the pheatmap R package (v0.7.7 or v1.0.2) was used to
966 scale and cluster only the rows, using a Euclidean distance metric and Ward clustering.

967

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

973

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

977 **Differentially Abundant miRs**

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

984

985 **Relationships Between Copy Number and miRNA Abundance**

986 In order to characterize how somatic copy number alterations (SCNA) influenced miRNA
987 abundance, MatrixEQTL v2.1.1¹⁰⁶ was used to calculate Spearman correlations between a)
988 normalized (RPM) abundance for the subset of pre-miRNAs (i.e. stemloops) that had an RPM of
989 at least 1.0 in at least 10 of the 178 tumor samples, and b) GISTIC2 real-valued (i.e. not
990 thresholded) SCNAs. SCNA data used Gencode v20 gene (miRNA) names, where 383 of the 476
991 stemloops selected by RPM above had Gencode names in the SCNA file, and another 28 had
992 overlapping genes with SCNA records (e.g. LPP for hsa-mir-28), allowing correlations to be
993 calculated for 411 of the RPM-selected stemloops. Correlations were thresholded at $FDR < 0.05$,
994 and for a subset of the miRNAs we generated both SCNA vs RPM scatterplots and full-
995 chromosome SCNA heatmap graphics using IGV 2.3.40. To generate a heatmap of global SCNA
996 vs. miR-based NMF unsupervised clustering, we imported the ‘seg’ data and NMF clustering
997 results into IGV v2.3.52, and ordered the samples to correspond to the 6-cluster miR-based

998 unsupervised NMF clustering heatmap. Samples were sorted in IGV by amplification at the
999 location of select miRNA in order to generate more focused whole-chromosome IGV graphics for
1000 a small number of miRNAs that had the strongest relationships with SCNA.

1001

1002 **Relationships Between Methylation and miRNA Abundance**

1003 An miRNA was considered to be epigenetically controlled if BH-corrected p-values were less than
1004 0.01 for both a) a Spearman correlation of miRNA abundance (RPM) to beta for probes in
1005 promoter regions associated with the miRNAs, and for b) a t-test of RPM between unmethylated
1006 ($\beta < 0.1$) and methylated ($\beta > 0.3$) samples (an ‘epigenetically-controlled pattern’).

1007

1008 **Relationships with EMT Scores**

1009 We identified miRNAs that have been associated with EMT⁶²⁻⁶⁶ and then calculated Spearman
1010 correlations between the EMT scores and RPMs for 5p and 3p mature strands for each of these
1011 miRNAs using MatrixEQTL and filtering by FDR<0.05. Heatmaps of miR abundance were
1012 generated for the miR-based unsupervised clusters for all samples (n=178) and squamous samples
1013 (n=144), sorting samples by EMT score within each unsupervised cluster and displaying only miRs
1014 whose correlations were larger than the median for each of the four cases. For *TGFBR2*, *CREBBP*,
1015 *EP300*, *SMAD4*, miR-200a, and miR-200b, we generated covariate tracks for alterations that
1016 included mutations and homozygous deletions downloaded from the cBio portal
1017 (www.cbioportal.org) and alterations in miR-200a and miR-200b (Methods and Supplemental
1018 Information S15).

1019 **miR Targeting**

1020 We assessed potential miRNA targeting for all 178 samples and then separately for the 144
1021 squamous samples by calculating miR-mRNA and miR-protein (RPPA) Spearman correlations
1022 with MatrixEQTL v2.1.1 using gene-level normalized abundance RNA-seq (RSEM) data and
1023 normalized RPPA data. Correlations were calculated with a p-value threshold of 0.05, and then
1024 the anti-correlations were filtered at $FDR < 0.05$. We extracted miR-gene pairs that corresponded
1025 to functional validation publications reported by miRTarBase v4.5²². For miR-RPPA anti-
1026 correlations, all gene names that were associated with each antibody were used. Results were
1027 displayed with Cytoscape v2.8.3.

1028

1029 **Relationships Between Endometrial and Cervical Tumor Samples**

1030 Analyses were performed to compare miR abundance profiles between this 178-sample cervical
1031 tumor set (CESC) and the TCGA cohort of 521 uterine corpus endometrial carcinomas (UCEC).
1032 First, we generated an unsupervised clustering solution using methods described above and
1033 annotated a selected clustering solution with the CESC vs. UCEC disease type, the CESC
1034 histological types, and the UCEC-like CESC samples (see Methods and Supplemental Information
1035 S5). miRs were then identified that were differentially abundant between UCEC and CESC
1036 samples with an unpaired two-class SAMseq v2.0 analyses with $FDR < 0.05$, as described above.

1037

1038

1039

1040 **Results**

1041 NMF unsupervised consensus clustering for 178 primary tumor samples suggested a six-cluster
1042 solution (Supplemental Fig. S12a, b). Median purities varied from 0.85 to 0.59 for the clusters
1043 (Supplemental Fig. S12c). Clusters were strongly associated with histology ($p=2.2e-17$), HPV
1044 clade ($p=0.0018$), and unsupervised clusters from other molecular platforms (Supplemental Fig.
1045 S12b). miR Clusters 5 ($n=30$) and 6 ($n=11$) separated the adenocarcinoma-enriched and HPV-
1046 negative samples into two subgroups; however, these samples were reported as a single cluster by
1047 iCluster (Adenocarcinoma cluster), PARADIGM (C2), and mRNA (C1). In contrast, the DNA
1048 methylation CIMP-high cluster was enriched only in miR Cluster 5. miRs that were differentially
1049 abundant between the clusters and also strongly abundant in at least one cluster (Supplemental Fig.
1050 S12d) included many that are known to be associated with cancer: miR-10a-5p, 21-5p, 22-3p, 143-
1051 3p, 182-5p, 203a, 205-5p, and 375. For example, for Clusters 5 and 6 noted above, both had
1052 relatively high miR-141-3p and miR-200a-3p, and relatively low miR-205-5p, while miR-10a-5p,
1053 21-5p, 30a, and 375 were more abundant in Cluster 5 than in Cluster 6. Cluster 2 had very high
1054 levels of miR-203a, Cluster 1 had high levels of miR-143-3p and low levels of miR-200 family
1055 members, Cluster 3 had the highest levels of the oncomiR miR-21-5p, and Cluster 4 had high
1056 levels of miR-205-5p.

1057

1058 For the five squamous miR-based clusters (Supplemental Fig. S13), many of the same miRs were
1059 differentially and highly abundant, such as miR-21-5p, 143-3p, 203a, and 205-5p (Supplemental

1060 Fig. S13d). Four of these five clusters corresponded to clusters from the n=178 six-cluster solution
1061 (Supplemental Fig. S12 and S13f).

1062

1063 There were no statistically significant differences between overall survival across the miR-based
1064 clusters for n=178 (Supplemental Fig. S12e; log-rank $p=0.34$) or for n=144 (Supplemental Fig.
1065 S13e; $p=0.13$).

1066

1067 **Differentially Abundant miRs**

1068 miRs that were differentially abundant between unsupervised clusters or other sample groups were
1069 identified by nonparametric unpaired two-class or multiclass analyses (Supplemental Table 14).
1070 miR-944¹⁰⁷ and 205-5p were strikingly more abundant in squamous than in adenocarcinoma
1071 samples, while miR-192-5p, 194-5p, and particularly 375 were less abundant (Supplemental Fig.
1072 S14b). Results were similar for HPV16-positive squamous vs. HPV16-positive adenocarcinoma
1073 samples (Supplemental Fig. S14c). For HPV16-positive squamous vs. HPV18-positive squamous,
1074 only miR-944 and 375 passed the $FDR<0.05$ threshold (Supplemental Fig. S14d). For HPV-
1075 positive vs. HPV-negative samples, miR-944 and the weakly abundant miR-767-5p and miR-105-
1076 5p were most strongly differential (Supplemental Fig. S14e).

1077

1078 **miRs Associated With Somatic Copy Number Alterations**

1079 While somatic copy number alterations were widespread, they were relatively weakly associated
1080 with miR clusters (Supplemental Fig. S15a). Of the miRNA stem-loops whose normalized RPM

1081 abundance was cis-correlated with SCNA, those with Spearman cis-correlations of at least 0.3 had
1082 low FDRs, and scatterplots were consistent with SCNA influencing miRNA abundance
1083 (Supplemental Fig. S15b, c, d). These miRNAs included a number that were involved in potential
1084 miR-gene targeting (Supplemental Figs. S17 and S18).

1085

1086 **Epigenetically Controlled miRNAs**

1087 The abundance of miR-10a, 17/18a/19a/20a, 141, 150, 152, and 205 appeared to be influenced by
1088 cis-DNA methylation, with miR-10a and 205 showing the clearest differences across miR-based
1089 clusters (Supplemental Fig. S16).

1090

1091 **Functionally Validated Potential miR-gene Targeting**

1092 We assessed potential miR targeting through miR-mRNA and miR-protein (RPPA) anti-
1093 correlations for all sample and squamous only sample cohorts (FDR<0.05, (Supplemental Table
1094 15)). Network graphics show the subset of high-confidence, FDR-thresholded anti-correlations
1095 that have been published as validated targets (Supplemental Figs. S17 and S18). The figures
1096 distinguish genes that are available only in mRNA data from those available in both mRNA and
1097 RPPA data. The figures also distinguish between anti-correlations identified with mRNA,
1098 nonphosphorylated proteins, and phosphorylated proteins. Many cancer-associated miRs were
1099 evident in the filtered anti-correlations. For example, a subnetwork involving miR-200-family
1100 miRs, the EMT-related transcription factors *ZEB1* and *ZEB2*, the Hippo effector *YAPI*, *ERBB2*,

1101 and *ERBB3* is presented in the all sample cohort. Fewer filtered targeting relationships are reported
1102 in the squamous sample cohort, some of which include *ZEB1*, *ZEB2*, and *ESR1*.

1103

1104 **Comparing Endometrial and Cervical Tumors**

1105 Unsupervised NMF consensus clustering of miR abundance profiles was used to compare 521
1106 TCGA endometrial tumor samples with the 178 cervical tumor samples. Clustering solutions
1107 appeared acceptable for between 9 and 15 clusters, which was the maximum assessed
1108 (Supplemental Fig. S19a). Details are reported for the 12-cluster solution (Supplemental Fig.
1109 S19b). In this solution, 9 clusters were exclusively or almost exclusively endometrial. Cluster 1
1110 was almost exclusively cervical, Cluster 3 was enriched for cervical samples, with endometrial
1111 samples generally less typical cluster members, and Cluster 8 was enriched for endometrial
1112 samples. Endometrial-like cervical cancer samples were distributed across four clusters. An
1113 unpaired two-class differential abundance analysis identified miR-944 and 205-5p as far more
1114 abundant in cervical than in endometrial tumor samples (Supplemental Fig. S19c and
1115 Supplemental Table 14).

1116

1117

1118

1119

1120

1121

1122 **S8. Reverse Phase Protein Array (RPPA) Analysis**

1123

1124 **RPPA Experiments and Data Processing**

1125 Frozen tumors were lysed using Precellys homogenization (Cayman Chemical, Ann Arbor,
1126 Michigan) and protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mM Hepes
1127 (pH 7.4), 150 mM NaCl, 1.5 mM MgCl₂, 1 mM EGTA, 100 mM NaF, 10 mM NaPPi, 10%
1128 glycerol, 1 mM phenylmethylsulfonyl fluoride, 1 mM Na₃VO₄, and aprotinin 10 µg/mL). RPPA
1129 was performed as described previously¹⁰⁸. Briefly, tumor lysate concentrations were adjusted to
1130 1 µg/µL as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates
1131 were manually serial-diluted in 5 two-fold dilutions with lysis buffer and printed on nitrocellulose-
1132 coated slides (Grace Bio-Labs) using an Aushon Biosystems 2470 arrayer (Billerica, MA). Slides
1133 were probed with 192 validated primary antibodies (Supplemental Table 17) followed by detection
1134 with appropriate secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG, or Rabbit anti-
1135 Goat IgG). The signal obtained was amplified using a Cytomation-catalyzed system of Avidin-
1136 Biotinylated Peroxidase (Vectastain Elite ABC kit from Vector Lab) binding to the secondary
1137 antibody and catalyzing Tyramide-Biotin (PerkinElmer) conjugation to form insoluble
1138 biotinylated phenols. Signals were visualized by a secondary streptavidin-conjugated HRP and
1139 DAB colorimetric reaction. The slides were scanned, analyzed, and quantified using Array-Pro
1140 Analyzer software (MediaCybernetics) to generate spot intensity (Level 1 data).
1141 SuperCurveGUI¹⁰⁹, which is available at
1142 <http://bioinformatics.mdanderson.org/Software/supercurve/>, was used to estimate relative protein

1143 levels (in log₂ scale). A fitted curve ("supercurve") was created with signal intensities on the Y-
1144 axis and relative log₂ amounts of each protein on the X-axis using a non-parametric, monotone
1145 increasing B-spline model¹⁰⁸. Raw spot intensity data were adjusted to correct spatial bias before
1146 model fitting using "control spots" arrayed across the slides¹¹⁰. A QC metric¹¹¹ was generated for
1147 each slide to determine slide quality and only slides with 0.8 on a 0-1 scale were used further. For
1148 replicate slides, the slide with the highest QC score was used for analysis (Level 2 data). Protein
1149 measurements were corrected for loading as described^{109,112} using median-centering across
1150 antibodies (Level 3 data). Seventeen samples with low protein levels were excluded from further
1151 analysis. In total, 192 antibodies and 155 samples were analyzed. Antibodies were selected to
1152 represent the breadth of cell signaling and repair pathways²³ conditioned on a strict validation
1153 process as previously described¹¹³. Antibodies are labeled as "validated" and "use with caution"
1154 based on degree of validation. Raw data (Level 1), SuperCurve nonparameteric model fitting data
1155 (Level 2), and protein loading corrected data (Level 3) were deposited at the DCC.

1156

1157 **Consensus Clustering**

1158 Consensus clustering was performed using an R package "ConsensusClusterPlus" to determine a
1159 robust number of sample clusters. Pearson correlation was used as a distance metric and Ward
1160 was used as inner and final linkage algorithm in the unsupervised hierarchical clustering analysis.
1161 Sample cluster number and membership were determined by stability evidence of 1000 resampling
1162 iterations. After consensus clustering analysis, 3 sample clusters were determined for all 155
1163 samples.

1164 **Silhouette Clustering**

1165 The consensus clusters of 155 samples were validated by Silhouette Clustering. Euclidean
1166 distance algorithm was used to compute the pairwise dissimilarities between samples. Out of 155
1167 samples, 115 whose Silhouette width was larger than 0.02 were retained as Silhouette Core
1168 samples for further analysis.

1169

1170 **Heatmap Generation**

1171 The Next Generation Clustered HeatMaps (NG-CHM) tool developed at the MD Anderson Cancer
1172 Center was used to generate heatmaps for the Level 3 RPPA data. Antibody clusters were
1173 determined by unsupervised hierarchical clustering in which Pearson correlation was used as a
1174 distance metric and Ward was linkage rule. For all samples, sample clusters were supervised by
1175 the consensus clusters. For the 115 Silhouette Core samples, sample clustering employed
1176 unsupervised hierarchical clustering using Pearson correlation as a distance metric and Ward as
1177 linkage rule.

1178

1179 **Statistical Analysis**

1180 Pathway scores were generated as described previously⁷ and the differences in pathway scores
1181 between RPPA clusters were evaluated by the non-parametric Kruskal-Wallis one-way ANOVA
1182 method. Correlation between RPPA clusters and other categorical variables were detected by Chi-
1183 Squared test, while correlations with continuous variables were examined using the non-
1184 parametric Kruskal-Wallis test. The significance of survival distributions between RPPA clusters

1185 was estimated by log-rank test and visualized with Kaplan-Meier survival curves. All statistical
1186 analyses were done using R (version 3.0.2).

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207 **S9. iCLUSTER Analysis**

1208

1209 **Data**

1210 Datasets used and transformations performed are described in Methods.

1211

1212 **iCLUSTER Method**

1213 Integrative clustering of RNA-seq, methylation, CNV, and mature-strand miRNA data was
1214 performed using R package “iCluster”²⁰. The method utilizes joint latent variable model within a
1215 likelihood framework with a lasso (L1) penalty in order to select the important features creating
1216 sparse solution. The tumor subtypes are modeled as unobserved latent variables which are
1217 simultaneously estimated from the multiple data types. Expectation Maximization (EM) algorithm
1218 is implemented for maximizing the penalized log-likelihood. Using the algorithm, posterior mean
1219 of the latent factor conditional to the data is estimated and then standard k-means clustering
1220 algorithm is used to draw inference on the cluster membership of the samples. Analyses were
1221 completed with all samples and then separately by histology (squamous and adenocarcinoma).

1222

1223 Optimum number of clusters k together with optimum sparseness parameter λ for L1 penalty is
1224 determined using the Proportion of Deviance (POD) method where the POD can be interpreted as
1225 the sum of the absolute differences between obtained cluster block structure and theoretical
1226 (perfect) block structure. Smaller POD indicates stronger cluster distinguishability.

1227

1228 In order to select an adequate number of features for the iCluster, analyses were carried out using
1229 500, 250, 100, and 50 most variable features from each dataset. There was a high degree of
1230 concordance among the resulting clustering assignments as measured by adjusted Rand Index.
1231 The results presented here are based on the 500 most variable features from each dataset.

1232

1233 Association analysis of clinical features and mutations with iCluster grouping was performed using
1234 Kruskal Wallis, Wilcoxon Rank-Sum, or Fisher's Exact tests. Differences in the survival of the
1235 subjects across the cluster groups were assessed using Kaplan-Meier analysis followed by a Log-
1236 Rank test. Heatmaps were made using the heatmap function in R package "NMF."

1237

1238 **Results**

1239 *All samples:* Integrative clustering was carried out using the 500 most variable features from each
1240 dataset. The integrative clustering identified three clusters consisting of 50, 86, and 42 samples.
1241 The Keratin-high cluster was entirely made up of squamous samples. The Keratin-low cluster was
1242 also enriched for squamous samples, while the Adenocarcinoma cluster contained most of the
1243 adenocarcinoma samples. Association analyses between the 14 significantly mutated genes
1244 (SMGs) identified by MutSig across the three clusters were carried out using Fisher's Exact test.
1245 *KRAS* ($p=9.74e-5$), *ERBB3* ($2.63e-3$), and *HLA-A* ($2.65e-2$) mutations were found to be
1246 significantly associated with clusters. *KRAS* mutations were not present in the Keratin-high cluster
1247 and *HLA-A* mutations were not present in the Adenocarcinoma cluster (Fig. 2). Further association
1248 analysis of mRNA-seq expression of these SMG genes across the 3 clusters were carried out using

1249 Kruskal Wallis test, with *NFE2L2* (4.56e-11), *TGFBR2* (4.62e-8), *ERBB3* (2.14e-7), *PIK3CA*
1250 (1.17e-4), *ARIDIA* (8.74e-4), and *KRAS* (3.19e-2) expression significantly associated with
1251 clusters.

1252

1253 Out of 178 total samples used for clustering, 112 samples had protein expression data. Association
1254 of protein expression with cluster groups was carried out using Kruskal Wallis test, and 54 proteins
1255 were significantly differentially expressed across the three clusters. Expression of Phospho-ERK
1256 (T202/Y204) (p=3.98e-2) that maps to the SMG *MAPK1* and HER2 (p=3.38e-3) that maps to
1257 *ERBB2* were found to be significantly associated with clusters. *APOBEC3A* (p=2.90e-14),
1258 *APOBEC3C* (p=1.16e-10), *APOBEC1* (p=3.20e-11), *APOBEC3B* (p=3.72e-2), and *APOBEC3G*
1259 (p=4.46e-2) gene expression were significantly different across the clusters using Kruskal Wallis
1260 Test. In addition, HPV16A vs. HPV16 non-A variants were significantly associated with the
1261 clusters (Fisher's Exact test p-value=0.002679).

1262

1263 ***Squamous cell carcinoma samples:*** Integrative clustering analysis on 144 samples of squamous
1264 histology identified 2 clusters with 97 and 47 samples. Association analysis with mutations in
1265 SMGs was carried out across the two clusters, with *KRAS* mutations being significantly associated
1266 with the clusters (p=0.01). mRNA-seq expression of the SMGs was assessed across the 2 clusters
1267 using Wilcoxon test, with *PIK3CA* (6.29e-6), *NFE2L2* (7.24e-6), *HLA-B* (1.07e-3), *TGFBR2*
1268 (2.82e-3), *EP300* (5.26e-3), *MAPK1* (5.47e-3), *HLA-A* (9.47e-3) and *FBXW7* (1.11e-2)
1269 significantly associated with the clusters.

1270

1271 Out of 144 total squamous samples, 92 samples had protein expression data. Association of protein
1272 expression with cluster groups was carried out using Wilcoxon test. Multiple proteins involved in
1273 MAPK, RTK, and Hippo pathway signaling were associated with the squamous clusters.
1274 *APOBEC3A* (p=3.09e-11), *APOBEC3C* (p=9.43e-5), *APOBEC3B* (p=1.82e-3), *APOBEC1*
1275 (p=5.13e-3), and *APOBEC3H* (p=2.73e-2) gene expression were significantly different across the
1276 clusters using Wilcoxon Test.

1277

1278 ***Adenocarcinoma samples:*** Integrative clustering analysis on 31 adenocarcinoma samples
1279 identified 2 clusters composed of 18 and 13 samples. Associations of gene mutations were carried
1280 out across the two clusters; however, mutations in the SMGs were not significantly associated with
1281 adenocarcinoma clusters. mRNA-seq expression of the SMGs were assessed across the 2 clusters
1282 using Wilcoxon test, with *ARID1A* expression significantly associated with clusters (p=2.76e-2).

1283

1284 Out of 31 total adenocarcinoma samples, 18 samples had protein expression data. Association of
1285 each protein expression with cluster groups was carried out using Wilcoxon test. Multiple proteins
1286 involved in metabolism and DNA damage repair were significantly associated with clusters. Gene
1287 expression of *APOBEC3D* (p=2.39e-4) and *APOBEC1* (p=4.94e-4) were significantly different
1288 across the clusters using Wilcoxon Test.

1289

1290

1291 **S10. PARADIGM Analysis**

1292

1293 **Data and Algorithm**

1294 The data and algorithm are described in Methods.

1295

1296 **Consensus Clustering of PARADIGM Inferred Pathway Activation**

1297 Consensus clustering based on the 3877 most varying features (i.e. IPLs with variance within the
1298 highest quartile) was used to identify subtypes implicated from shared patterns of pathway
1299 inference. Consensus clustering was implemented with the ConsensusClusterPlus package in R¹¹⁴.

1300 Specifically, median-centered IPLs were used to compute the squared Euclidean distance between
1301 samples, and this metric was used as the input to the ConsensusClusterPlus algorithm.

1302 Hierarchical clustering was performed using the Ward's minimum variance method (i.e. ward
1303 inner linkage option) and 80% subsampling was performed over 1000 iterations, with the final
1304 consensus matrix clustered using average linkage. The number of clusters was selected by
1305 considering the relative change in the area under the empirical cumulative distribution function

1306 (CDF) curve as well as the average pairwise item-consensus within consensus clusters. We
1307 selected k=4 as further separation provides minimal change and decreases the within-cluster
1308 consensus. Heatmap display of the top varying IPLs was generated using the heatmap.plus

1309 package in R. Differences in overall survival (OS) between PARADIGM clusters were assessed
1310 by the log-rank test, and the chi-square test was used to evaluate associations with clinical

1311 parameters (histology and HPV clade) and single platform subtypes (mRNA, copy number,
1312 methylation, miRNA, and RPPA clusters).

1313

1314 Pathway biomarkers of each PARADIGM cluster (vs. all others) were identified using the t-test
1315 and Wilcoxon Rank-Sum test with Benjamini-Hochberg (BH) false discovery rate (FDR)
1316 correction. Only features deemed significant (FDR corrected $p < 0.05$) by both tests and showing
1317 an absolute difference in group means > 0.05 were considered. Interconnectivity between these
1318 pathway biomarkers within the PARADIGM SuperPathway was assessed, and regulatory hubs
1319 with ≥ 10 differentially activated downstream targets were selected and displayed in a heatmap.

1320

1321 **Pathway Biomarkers Differentiating Squamous Carcinomas and Adenocarcinomas**

1322 IPLs differentially activated between squamous carcinomas (n=144) and adenocarcinomas (n=31)
1323 were identified using the t-test and Wilcoxon Rank-Sum test with BH FDR correction. Only
1324 features deemed significant (FDR corrected $p < 0.05$) by both tests and showing an absolute
1325 difference in group means > 0.05 were selected. Differentially activated IPLs were then filtered
1326 by connectivity within the SuperPathway, such that only interconnected features via regulatory
1327 interactions were retained. Pathway constituents of the PARADIGM SuperPathway enriched
1328 among these selected features were assessed using the EASE score with BH FDR correction, and
1329 subnetworks were constructed to identify regulatory hubs with ≥ 10 outgoing regulatory edges and
1330 visualized using Cytoscape.

1331

1332 Interconnected complexes and features (by any edge type) showing differential activation between
1333 squamous and adenocarcinomas within the FGFR3 network neighborhood were visualized in
1334 Cytoscape. In addition, the mRNA expression levels of FGFR1 and FGFR3 were compared using
1335 Spearman rank correlation and differences in expression of these genes in squamous vs.
1336 adenocarcinomas were visualized using box plots.

1337

1338 In order to illustrate the difference in p63 inferred pathway activation between squamous cell
1339 carcinomas and adenocarcinomas, a heatmap of the scaled (mean 0 and standard deviation 1) p63
1340 PARADIGM inferred activity, scaled log₂-transformed mRNA expression, and GISTIC
1341 thresholded copy number levels ordered by sample histology was constructed using the
1342 heatmap.plus package in R. The log₁₀-transformed expression of the top two differential miRNAs
1343 between squamous vs. adenocarcinoma - miR944 and miR205 - were also scaled and included in
1344 the heatmap, and the expression of these miRNAs was compared to p63 mRNA expression levels
1345 using Pearson correlation.

1346

1347 **Pathway Biomarkers Associated With HPV Status**

1348 IPLs differentially activated between HPV Clade A9 (n=120) vs. Clade A7 (n=45) were identified
1349 using the t-test and Wilcoxon Rank-Sum test with BH FDR correction. Only features deemed
1350 significant (FDR corrected p<0.05) by both tests and that showed an absolute difference in group
1351 means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity within
1352 the SuperPathway, such that only interconnected features (at least 1 interaction of any kind) were

1353 retained. Subnetworks linked through regulatory (activation or inhibition) interactions were
1354 constructed and visualized using Cytoscape, and constituent pathways of the PARADIGM
1355 SuperPathway enriched within these subnetworks were assessed using the EASE score with BH
1356 FDR correction. This analysis was also performed restricted to the squamous histology subtype
1357 (A9: n=103, A7 n=35). A similar analysis was performed to identify pathway biomarkers
1358 distinguishing HPV negative (n=9) from HPV positive (n=169) cases.

1359

1360 **Results**

1361 Consensus clustering using the top varying PARADIGM inferred pathway levels (IPLs) yields 4
1362 subtypes with characteristic patterns of pathway activation (Supplemental Fig. S48). Of note, 29
1363 of 31 adenocarcinomas are clustered together in PARADIGM C2, which also contains 7 of the 9
1364 HPV-negative cases. In addition to associations with histology, PARADIGM subtypes also show
1365 significant associations with HPV clade as well as other single platform subtypes. Highest inferred
1366 activation of FOXA2 and XBP1-2 pathways is observed within the adenocarcinoma-enriched
1367 PARADIGM C2. Key pathway features distinguishing PARADIGM cluster C4 from non-C4
1368 cases include highest relative inferred activities of pathways involving DNA damage, MYB, and
1369 IL-12. PARADIGM cluster C3 is associated with highest inferred FOXM1 and MYC pathway
1370 activation, while the remaining PARADIGM cluster C1 samples show highest inferred activation
1371 of HIF1A, STAT6, p53, p63, p73, ARF2, and ERK signaling.

1372

1373 Of the 4692 PARADIGM IPLs identified as differentially activated between adenocarcinomas and
1374 squamous cell carcinomas, 1098 are connected through regulatory interactions (activation or
1375 inhibition) (Extended Data Fig. 10). Pathway enrichment and subnetwork analysis of the
1376 interconnected differential pathway features implicates higher activation of FOXA1/ER and
1377 FOXA2 pathways in adenocarcinomas. In contrast, key distinguishing features of squamous
1378 carcinomas include higher inferred activation of p53, p63, p73, AP-1, MYC, HIF1A, and MAPK
1379 signaling. Interestingly, inferred p63 activation and to a greater extent p63 mRNA expression
1380 levels show significant correlations with the two most differentially abundant miRNAs between
1381 squamous and adenocarcinomas: miR-944 and miR-205. Also of note, FGFR3 appears to have
1382 higher inferred activity in squamous carcinoma, likely attributable to higher mRNA expression
1383 levels within this histological subtype. Paradoxically, FGFR1 mRNA levels, which show a modest
1384 but significant negative correlation with FGFR3 expression, appear higher in adenocarcinomas.

1385
1386 A comparison of PARADIGM inferred pathway activation between Clade A7 vs. Clade A9 HPV
1387 positive samples identifies higher inferred activation of p53 and p63 signaling and lower FOXA1
1388 signaling in the Clade A9 infected cases. These significant differences are retained when the
1389 analysis is restricted to the squamous subtype (Fig. 5a). Consistent with expectations, inferred
1390 activation of NF-kB signaling appears lower in HPV-negative relative to HPV-positive samples.
1391 Interestingly, lower inferred activity of p53 and MAPK3 signaling is also observed.

1392

1393

1394 **S11. APOBEC Mutagenesis Analysis**

1395

1396 **Data Deposition**

1397 Complete output of APOBEC mutagenesis analysis used for this paper in the format of Broad
1398 Institute GDAC Firehose is in the APOBEC_CESC_res3_192.7z folder placed under controlled
1399 access at: <https://tcga-data-secure.nci.nih.gov/tcgafiles/tcgajamboree/CESC/APOBEC/>.

1400

1401 In order to navigate through data all files should stay in the same folder. The
1402 “192_genome.wustl.edu_CESC.IlluminaGA_DNASeq_curated.Level_2.1.0.0.somatic.maf_sorte
1403 d_report.html” Nozzle output file provides detailed legends and annotated links to all data files.
1404 A partial set of files containing Nozzle output, graphics summaries of analysis, and the most
1405 important data files are provided in the open access APOBEC_output.zip file on the TCGA
1406 Publication Page Portal.

1407

1408 **Methods**

1409 The exome-wide prevalence of the APOBEC mutagenesis signature and the enrichment of this
1410 signature over its presence expected for random mutagenesis were evaluated as described
1411 previously¹⁵ with some additions (see Methods). On top of previously described output, several
1412 other parameters were calculated and annotations added that characterize the prevalence of the
1413 APOBEC mutagenesis pattern in a sample and/or that are useful for downstream analyses and
1414 comparisons. The main new parameter used in this study was the minimum estimate of the number

1415 of APOBEC-induced mutations in a sample, which is given the name
1416 “APOBEC_MutLoad_MinEstimate.” Values were calculated as described in Methods and are
1417 rounded to the nearest whole number.

1418

1419 The complete description of data files and columns in data tables are in readme files within the
1420 analysis output APOBEC_CESC_res3_192.7z folder under controlled access and within the open
1421 access APOBEC_output.zip file. The values of “APOBEC_MutLoad_MinEstimate” and category
1422 assignments for each sample are also presented in Supplemental Table 1.

1423

1424 **Results**

1425 Prior research has identified a stringent mutation signature tCw→tTw or tCw→tGw (mutated
1426 nucleotide is capitalized; w=A or T) characteristic of mutagenesis by a subclass of APOBEC
1427 cytidine deaminases abundant in many samples of cervical and other cancer types^{8,12,13,15}. In this
1428 study, 150 out of 192 exomes displayed statistically significant ($q < 0.05$) enrichment (up to 6-fold)
1429 with this signature. The signature was carried by 46% of all mutations in the dataset, approaching
1430 70% in some exomes. Even the minimum estimate accounting for the random mutagenesis
1431 resulting in a fraction of APOBEC signature mutations indicated that up to 1500 mutations in an
1432 exome can be caused by APOBECs (Supplemental Fig. S26 and APOBEC_output.zip). APOBEC
1433 mutation load strongly correlated with the total number of mutations in a sample (Extended Data
1434 Fig. 2h), suggesting that APOBEC mutagenesis is the major source of mutations in cervical
1435 cancers. HPV infection, which has been previously linked with increased APOBEC mutagenesis

1436 in head and neck cancers¹⁴, was also correlated with a pattern of APOBEC mutagenesis in cervical
1437 cancer samples (Supplemental Fig. S27). The cause of mutagenesis may be due to high expression
1438 of *APOBEC3* genes as a result of HPV at some point during (or before) cancer development, since
1439 transcription of APOBECs is known to be induced by factors triggering the innate immune
1440 response¹¹⁵. Indeed, expression of *APOBEC3A* showed the strongest positive correlation with
1441 mutagenesis and *APOBEC3B* showed overall high expression in cancers of the dataset
1442 (Supplemental Fig. S28). Mutagenesis could also be a consequence of DNA damage response
1443 (DDR) caused by HPV¹¹⁶, resulting in increased formation of single-stranded (ss) DNA – the
1444 exclusive substrate for APOBEC cytidine deaminases. Many mutations in genes with a potential
1445 role in the initiation and/or progression of cervical cancer carried the APOBEC mutagenesis
1446 signature, with *PIK3CA* harboring the most (Extended Data Fig. 2g) similar to observations in
1447 head and neck cancers¹⁴.

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457 **S12. EMT mRNA Score Analysis**

1458

1459 **Methods**

1460 The EMT score was computed as previously described^{10,21}. Briefly, the EMT score was the value
1461 resulting from the difference between the average expression of mesenchymal (M) genes minus
1462 the average expression of epithelial (E) genes. All NA values were removed from the calculation.
1463 Two-sample t-test and ANOVA were applied to each comparison accordingly. A Cox proportional
1464 hazards model was applied to assess whether the EMT score was associated with overall survival.
1465 Kaplan-Meier plots (Log-rank test) were used to display the difference between groups (the
1466 median value of EMT score of samples).

1467

1468 **Results**

1469 EMT scores were significantly higher in UCEC-like cancers (two sample t-test, $p=0.048$)
1470 (Supplemental Fig. S29b). Patients with higher EMT scores had worse overall survival ($p=0.0221$,
1471 log-rank test between top and bottom median EMT score patient groups) (Supplemental Fig.
1472 S29a). EMT scores were associated with the subtypes defined by different molecular platforms,
1473 including methylation CIMP ($p=0.024$), iCluster (0.003), miRNA ($p < 0.001$), mRNA ($p=0.003$),
1474 and PARADIGM ($p=0.005$), which suggests the association between EMT score and global
1475 molecular alterations at different levels (Supplemental Fig. S30).

1476

1477

1478 **S13. Functional Epigenetic Module (FEM) Analysis**

1479

1480 **FEM Algorithm**

1481 The Functional Epigenetic Module (FEM) algorithm³⁹ was used to identify potentially disrupted
1482 signaling pathways between groups. FEM represents a tool for the integrative analysis of DNA
1483 methylation and gene expression data that uses protein-protein-interaction (PPI) networks¹¹⁷ as the
1484 backbone for identifying subnetworks of genes that are epigenetically and functionally
1485 dysregulated based on a phenotype of interest. This methodology consists of two main parts: (i)
1486 computation of edge weights for connected genes in the PPI network where the weights are a
1487 composite measure of each gene's strength of association between both gene expression and DNA
1488 methylation and the phenotype of interest, and (ii) identification of subnetworks of genes where
1489 the average weight density is significantly larger than the rest of the network.

1490

1491 We began by subsetting the data to consist of the set of genes (G) that overlapped between the
1492 gene expression data, DNA methylation data, and genes represented in the PPI network. We then
1493 summarized DNA methylation information at the gene level by computing the average methylation
1494 of CpG sites mapping to within 200 bp of the transcription start site (TSS200). If there were no
1495 probes mapping to within 200 bp of the transcription start site, the average methylation of CpGs
1496 mapping to within the 1st exon of the gene was computed. If there were no probes mapping to
1497 within the 1st exon of the gene, the average methylation of CpGs mapping to within 1500 bp of
1498 the TSS (TSS1500) was computed. We next calculated the test-statistics, $t_g^{(R)}$ and

1499 $t_g^{(D)}$ $g=1,2,\dots,G$, obtained from testing the association between both gene expression and DNA
1500 methylation with the phenotype of interest for each of the G genes. A composite test-statistic for
1501 each gene t_g , $g=1,2,\dots,G$ was then computed. For genes exhibiting anti-correlation between
1502 gene expression and DNA methylation (i.e. $sign(t_g^{(R)}) \neq sign(t_g^{(D)})$), composite test-statistics were
1503 taken to be the absolute difference of the DNA methylation- and gene expression-based test-
1504 statistics (i.e. $t_g = |t_g^{(D)} - t_g^{(R)}|$); otherwise, $t_g = 0$ if $sign(t_g^{(R)}) = sign(t_g^{(D)})$. Weights between
1505 connected genes, gene g and gene h , in the PPI were taken to be the average of the composite test-
1506 statistics for those two genes (i.e. $w_{gh} = \frac{1}{2}(t_g + t_h)$). Lastly, the PPI network was scanned using a
1507 version of the spin-glass algorithm¹¹⁸ to identify subnetworks where the average weight density of
1508 connected genes was significantly larger than the rest of the network. The output of the FEM
1509 methodology is a series of subnetworks whose average weight density is statistically significantly
1510 greater than would be expected by chance.

1511

1512 The analyses described above were carried out using the Bioconductor package ‘FEM’ within the
1513 R statistical programming language.

1514

1515 **Results**

1516 In an attempt to understand the implications of HPV subtype on the underlying biology of cervical
1517 tumors, we considered several different applications of the FEM methodology to the cervical

1518 cancer data. Specifically, FEM was used to identify disrupted subnetworks between HPV clade
1519 A7 and A9 tumors and HPV-positive and -negative tumors. Identification of disrupted
1520 subnetworks between these groups was carried out using all Core Set samples (n = 178) and within
1521 squamous cell carcinomas (n = 144). In addition, we also examined disrupted subnetworks
1522 between HPV A7 and A9 adenocarcinoma tumors (n = 31). There were a total of $G = 6,730$ genes
1523 that overlapped between the DNA methylation data, gene expression data, and the PPI network.
1524 The total space for identifying disrupted subnetworks therefore consisted of a PPI network spanned
1525 by the 6,730 overlapping genes and the interactions between them.

1526

1527 **Identification of disrupted subnetworks between HPV-positive and HPV-negative tumors**

1528 Although only 9 out of 178 cervical tumors were HPV-negative, our analysis revealed 13
1529 statistically significant subnetworks ($p < 0.05$) when FEM was applied to the data consisting of all
1530 cervical histological subtypes (Supplemental Fig. S31 and Supplemental Table 19: Tab S1). The
1531 size of these 13 subnetworks ranged from as small as 10 genes to 44 genes. Interestingly, 3 out of
1532 the 13 identified subnetworks were centered around genes belonging to the Fibroblast Growth
1533 Family (FGF), specifically *FGF3*, *FGF4*, and *FGFR1*. Each of these genes showed statistically
1534 significant increased promoter DNA methylation ($p = 1.3e-6$, $6.2e-4$, and $3.8e-5$, respectively) and
1535 reduced expression ($p = 3.4e-9$, $1.6e-11$, and $1.2e-6$, respectively) in HPV-positive compared with
1536 HPV-negative cervical tumors. These findings are in agreement with recent data demonstrating
1537 that HPV16 E6/E7 infection (the predominant HPV subtype in these samples) partially represses

1538 the proliferation, but not the invasive potential, of cervical cancer cells stimulated by FGF2 or
1539 FGF4¹¹⁹.

1540

1541 Restricting analysis to only the squamous cell carcinomas (n = 144), 12 statistically significant
1542 subnetworks between HPV-positive (n = 140) and HPV-negative (n = 4) tumors were identified
1543 (Supplemental Table 19: Tab S2). Similar to the results obtained from fitting FEM using all
1544 cervical histologies, 2 out of the 14 statistically significant subnetworks were centered around FGF
1545 genes, specifically *FGF3* and *FGF4*.

1546

1547 To see if the disrupted subnetworks between HPV-positive and HPV-negative cervical squamous
1548 cell carcinomas were specific to cervical cancer, we next applied the FEM methodology to the
1549 HNSC dataset. In a similar manner, FEM was applied to the HNSC dataset for identifying
1550 disrupted subnetworks between HPV positive (n = 36) and HPV negative (n = 243) HNSC tumors.
1551 This analysis revealed 11 statistically significant subnetworks, which ranged in size from as small
1552 as 18 genes to as large as 62 genes (Supplemental Table 19: Tab S3). Although these 11
1553 subnetworks were largely distinct from the 12 statistically significant subnetworks between HPV
1554 positive and HPV negative cervical squamous cell tumors, there was one common subnetwork
1555 centered around Forkhead Box A2 (*FOXA2*) (Supplemental Table 19: Tabs S2, S3). Interestingly,
1556 *FOXA2* showed significantly increased promoter methylation and decreased expression in HPV
1557 positive compared to HPV negative cases in both the HNSC tumors and squamous cell cervical
1558 tumors (Supplemental Fig. S32). It is also worth noting that many of the genes contained in the

1559 *FOXA2* subnetwork showed consistent relationships between DNA methylation/gene expression
1560 and HPV status between the HNSC and squamous cell cervical tumors. These findings may
1561 suggest a common pathway(s) by which HPV exerts its effects on tumorigenesis.

1562

1563 **Identification of disrupted subnetworks between HPV A7 and A9 tumors**

1564 We also identified disrupted subnetworks between samples infected with HPV A7 vs. A9 clades.
1565 Applying FEM to the data consisting of all cervical histological subtypes, 8 statistically significant
1566 subnetworks were identified between HPV A7 (n = 45) and HPV A9 (n = 120) tumors
1567 (Supplemental Table 19: Tab S4). Restricting analysis to only the squamous cell cervical
1568 carcinomas (n = 136) revealed 7 statistically significant subnetworks (Supplemental Table 19: Tab
1569 S5). In the analysis restricted to non-squamous cell cervical carcinomas (n = 27), 4 statistically
1570 significant subnetworks between HPV A7 (n = 8) and HPV A9 (n = 19) tumors were identified
1571 (Supplemental Table 19: Tab S6).

1572

1573

1574

1575

1576

1577

1578

1579

1580 **S14. Immune Response Gene Analysis**

1581

1582 **Immune Response Gene Expression Analysis**

1583 The Core Set (144 squamous cell carcinomas (SCCs), 31 adenocarcinomas (ACs) and 3
1584 adenosquamous carcinomas) samples were used in this analysis and a total of 372 genes were
1585 selected based on GO 0006954 and 0006955 annotations. The gene symbols from GO selection
1586 were merged with the mRNA-seq matrix (Supplemental Table 20).

1587

1588 **Clustering Analysis**

1589 Consensus clustering (CC) analysis was performed based on the top 300 most variable genes
1590 filtered by median absolute deviation using the ConsensusClusterPlus package in R. The gene
1591 count numbers were log-transformed and median-centered. The agglomerative hierarchical
1592 clustering algorithm using Pearson correlation distance was performed using 80% item resampling
1593 (pItem), 100% gene resampling (pFeature), a maximum of 12 cluster counts (maxk), 1,000
1594 resampling (reps), and a random number of seed. The total number of clusters (k) was determined
1595 by the inspection of consensus cumulative distribution function (CDF) curves shape, and the
1596 relative change in area under the CDFs curve¹²⁰.

1597

1598 **Prognostic Cluster Analysis**

1599 An *ExpressionSet* class was designed with the TCGA mRNA-seq normalized matrix for gene
1600 expression analysis (assaydata), which included the 372 immune and inflammatory response genes

1601 and the *AnnotatedDataFrame* based on the vital status presented in Supplemental Table 1 using
1602 the Biobase package in R. The areas under the curves (AUCs) were calculated based on each gene
1603 expression and the living or deceased outcomes using the rowpAUCs function (genefilter package
1604 in R). Genes that failed to accurately predict survival ($AUC \leq 0.61$) were excluded¹²¹. The
1605 consensus clustering analysis was performed based on the selected genes as described above.

1606

1607 To analyze the association of immune cytolytic activity (CYT) score with prognostic clusters and
1608 overall survival, the geometric means of *GZMA* and *PFRI* genes in SCC samples were estimated
1609 (Supplemental Table 21)¹⁸.

1610

1611 The expression of 372 genes was compared between immune response and prognostic clusters.
1612 After the estimation of the dispersion for each gene using the “estimateTagwiseDisp” function,
1613 differentially expressed (DE) genes were identified by the exact test using edgeR package. Genes
1614 with log fold-change ($\log_{2}FC$) > 1.0 and false discovery rate (FDR) adjusted p-value < 0.05 were
1615 considered.

1616

1617 **Gene Set Enrichment Analysis (GSEA)**

1618 GSEA was performed based on the 372 immune gene expression matrix using the GSEA software
1619 and the Molecular Signature Database (MSigDB) REACTOME-c2.cp.reactome.v4.0.symbols.gmt
1620 (<http://www.broad.mit.edu/gsea/>). One thousand total permutations were used, and SCC versus
1621 AC and prognostic cluster C1 versus C2 were used as phenotype labels. The gene_set profile was

1622 used as the permutation type. Cytoscape software was used to create the Enrichment map. The
1623 WEB-based GENE SeT AnaLysis Toolkit (gestalt) was used to analyze common gene pathways
1624 into each gene cluster (<http://bioinfo.vanderbilt.edu/webgestalt/>).

1625

1626 **Survival Analysis**

1627 The survival analyses for immune response clusters and prognostic clusters were carried out using
1628 Kaplan-Meier curve and Cox proportional-hazards regression model in Rstudio.

1629

1630 **Results**

1631 Consensus clustering analysis identified five immune response clusters, with most ACs (n= 29)
1632 clustering together in cluster 5. Two adenosquamous samples cluster in C5 and one in C4. Among
1633 the 372 immune response genes analyzed, 83 were differentially expressed in C5 samples versus
1634 all other samples (Supplemental Table 22). Four gene clusters were differentially expressed in C5
1635 when compared to all other samples (Supplemental Fig. S33). Gene cluster 1 (blue) contains 9
1636 downregulated genes in C5 (*IL1A*, *IL1RAP*, *LTB4R*, *S100A8*, *S100A9*, *S100A12*, *GPR68*, *SPINK5*
1637 *and KRT1*). *IL1A* and *IL1RAP* are involved in IL1 signaling, and *S100A8* / *S100A9* are involved
1638 in endogenous toll-like receptor signaling. Cluster 2 (red) includes 3 downregulated genes in C5
1639 (*CD274*, *PDCD1LG2*, *AIM2*). *CD274* and *PDCD1LG2* are involved in adaptive immune response
1640 and costimulation by CD28 family signaling. Cluster 3 (magenta) includes 4 downregulated genes
1641 in C5 (*APOL3*, *CXCL9*, *CXCL10* and *CXCL11*). *CXCL9*, *CXCL10* and *CXCL11* are involved in
1642 CXCR3-mediated signaling events. *CD274* and *PDCD1LG2* genes encode PDL1 ligands PDL1

1643 and PDL2 protein, respectively. PDL1 is expressed in various solid tumors including squamous
1644 cell carcinomas of the lung, esophagus, and head and neck¹²². These proteins suppress T-cell
1645 effector function including the cytotoxic activity and their expression is induced by inflammatory
1646 cytokines¹²³. The use of PD1 immune blockage has resulted in long-term response in a subgroup
1647 of patients with lung cancer and melanoma^{124,125}. The loss of AIM2 protein (absent in melanoma
1648 2 protein) expression has been demonstrated as a prognostic marker in colorectal cancer¹²⁶, and is
1649 associated with metastatic dissemination in melanoma and cutaneous squamous cell
1650 carcinomas¹²⁷. dsDNA viruses are sensed by AIM2, triggering inflammasome formation and IL1B
1651 release. This mechanism is a key activator of innate and adaptive immune response¹²⁸. A recent
1652 study demonstrated that the AIM2 inflammasome is activated by HPV16 in keratinocytes¹²⁹.
1653 *CXCL9*, *CXCL10* and, *CXCL11* genes encode CXCR3 ligand cytokines known as angiostatic CXC
1654 chemokines¹³⁰, and are potent angiogenesis inhibitors linked to cell-mediated immunity. Cluster
1655 4 (pale green) contains upregulated genes in C5 (*ADORA1*, *DPP4*, *NFATC4*, *CRHR1*, *TCF7*,
1656 *FCGRT* and *CCR9*). *ADORA1*, *CRHR1* and *CCR9* are involved in G protein-coupled receptor
1657 binding. Cluster 5 (yellow) has 12 upregulated genes (*GPR44*, *SIGIRR*, *XBPI*, *ELF3*, *HLA-J*,
1658 *CHRNA7*, *HDAC9*, *SKAP1*, *CCBP2*, *MNX1*, *CHST4* and *ALOX15*) that are not enriched in a
1659 common pathway.

1660
1661 GSEA identified four significantly enriched Reactome pathways in SCCs compared with ACs.
1662 The “immune response” pathway is enriched by the “innate immune system” node and the
1663 “adaptive immune system” and its subfamily “costimulation by the CD28 family” nodes. There

1664 are 42 genes enriched and all of them are overexpressed in SCCs compared with ACs. The *CD274*,
1665 *PDCD1LG2*, *PDCD1*, *CD80*, *CD86*, and *CTLA4* genes are in the “costimulation by the CD28
1666 family node” and are involved in T-cell activation. Other genes involved in T-cell activation that
1667 are highly expressed in SCCs include *CD8A*, *CD28*, *GZMA*, and *PRF1*. The median CYT score
1668 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).
1669 Together, these results suggest that the adaptive immune response is repressed in ACs compared
1670 with SCCs. Adaptive immune response and T-cell modulation have been reported as promising
1671 therapies in human cancer^{131,132}. Our data suggest that the use of immune co-stimulatory
1672 molecules may be a potential therapy for cervical ACs. Based on the clustering analysis, there is
1673 a subset of SCCs with a low immunogenic profile similar to ACs. In order to determine whether
1674 immune gene expression can select groups of patients with distinct prognosis, a prognostic
1675 clustering algorithm was developed.

1676
1677 **Prognostic clusters:** ROC analysis was used to identify groups of patients with diverse prognosis
1678 in cervical carcinomas. The ROC analysis identified 47 genes with AUC > 0.61 (Supplemental
1679 Table 23). Using this set of genes, cervical carcinomas can be clustered into two different
1680 expression subtypes (Supplemental Fig. S34a), with C2 samples having worse prognosis compared
1681 with C1 samples (C1 versus C2, HR= 2.9; p= 0.002) (Supplemental Fig. S34b).

1682 Of the 259 differentially expressed genes between prognostic cluster C1 and C2 samples
1683 (Supplemental Table 24), 57 enriched genes were identified by GSEA (Supplemental Table 25).
1684 The main nodes enriched in C1 are “immune signaling,” “TCR and downstream TCR signaling

1685 pathways,” “adaptive immune system,” and “costimulation by CD28 family.” *PDL1/2*, *PDC1*,
1686 *CD86/CTLA4*, *CD40/CD40LG*, and *CD80/CD28* genes are overexpressed in prognostic cluster C1
1687 tumors, indicating that costimulatory and coinhibitory receptors are modulating T-cell activity.
1688 The enriched genes in prognostic cluster C2 samples are all associated to signaling by ILs (*IL1A*,
1689 *IL1R2*, *IL6*, *IL6ST*, *TRAF6*, *RIPK2* and *MAP3K7*).

1690

1691 When analyzing the association between CYT score and the prognostic clusters, a significantly
1692 higher CYT score was observed in C1 samples compared with samples in C2 cluster samples
1693 (Supplemental Fig. S34c). The linear regression model demonstrated that all genes associated
1694 with T-cell immune synapses are correlated with CYT, especially *PDCD1* ($r^2= 0.57$), *CTLA4* ($r^2=$
1695 0.57), *LAG3* ($r^2= 0.57$) and *CD86* ($r^2= 0.45$) (Supplemental Table 26).

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706 **S15. MEMo Analysis**

1707

1708 **MEMo Analysis**

1709 miRNA binary alteration calls and MEMo analysis are described in Methods.

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727 **S16. Mitochondrial DNA Analysis**

1728

1729 **Analysis Methods**

1730 Aligned BAM files from whole genome sequencing (WGS) analysis were used to extract reads
1731 aligned to mitochondria and GATK¹³³. Unified Genotyper was used to call SNVs and indels.
1732 Variants detected in the tumor but not in the corresponding normal were called as somatic. Somatic
1733 events were annotated using the MITOMAP database (<http://www.mitomap.org/MITOMAP>).
1734 Primary tumors and blood samples showed slightly different number of mitochondria, with their
1735 medians being 59 and 80, respectively. By WGS, the coverage on the MT genome was sufficient
1736 to call somatic mutations, whereas in whole exome sequencing (WES) these regions were not
1737 selected for. However, when calling mitochondrial mutations on samples using WES data, we
1738 were able to recall 71% of variants made using WGS data.

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748 **S17. RNA Splicing Analysis**

1749

1750 **Detecting RNA Splicing Events**

1751 SpliceSeq¹³⁴ was used to analyze RNA-seq data for transcript splicing variation. SpliceSeq
1752 aligns reads to splice graphs representing all protein coding isoforms of human genes in
1753 Ensembl. Percent spliced in (PSI) values are generated for each potential splice event for all
1754 samples and all genes. The type of splice events detected include exon skip (ES), retained intron
1755 (RI), alternate donor (AD), alternate acceptor (AA), mutually exclusive exon (ME), alternate
1756 promoter (AP), and alternate terminator (AT). PSI is the ratio of normalized read counts
1757 indicating the inclusion path vs. the total covering a splice event (Supplemental Fig. S37). For
1758 further details on SpliceSeq methods, see:
1759 <http://bioinformatics.mdanderson.org/main/SpliceSeqV2:Methods>.

1760

1761 To evaluate changes in splicing patterns across CESC samples, a subset of splice events
1762 demonstrating variation across tumor samples was selected. The splice event selection criterion
1763 was: 1) Minimum average expression RPKM > 1.5; 2) PSI values for 95% of the samples; 3)
1764 occurrence in a highly expressed portion of the transcript (magnitude > .3); and 4) PSI standard
1765 deviation across samples of > .25. For genes with more than one splice event meeting this criterion,
1766 the splice event with the strongest average read coverages was selected. The resulting 219 splice
1767 events represent those with the strongest differential splicing behavior across the Core Set of

1768 samples. The full set of differential splicing event PSI values is provided in Supplemental Table
1769 28.

1770

1771 **Results**

1772 The PSI values of selected splice events were mean-centered across samples and used to create a
1773 hierarchical clustered heatmap of sample vs PSI (Distance Metric = Correlation; Agglomeration
1774 Method = Ward). The samples clustered into 3 clusters that were further investigated in
1775 downstream analysis (Supplemental Fig. S38). Fisher's Exact tests were performed to evaluate
1776 similarity between splicing clusters and clinical data/clusters from other platforms. Splicing
1777 cluster 2 (orange) contains the majority (24 of 31, Fisher's p-value < 0.001) of adenocarcinoma
1778 samples, and therefore overlaps with many of the other adenocarcinoma-enriched platform clusters
1779 (30 of 42 Adenocarcinoma iCluster samples, 32 of 47 mRNA C1 samples, 22 of 30 miRNA C5
1780 samples, and 31 of 45 PARADIGM C2 samples).

1781

1782 Splicing clusters 1 and 3 both contain predominantly squamous samples but Cluster 3 contains a
1783 smaller subset of squamous samples that displays a strikingly different pattern of alternative
1784 splicing. In general, Cluster 3 does not have strong associations with clusters identified by the
1785 other platforms so this appears to be a unique subset of squamous samples identified by splicing
1786 analysis. The only exception is an association with PARADIGM C4 (20 of 28 members are in
1787 Splicing Cluster 3; p<0.001). Cluster 3 had no significant association with clinical annotations
1788 with the exception of vital status. Only two of the 26 patients who died are in Cluster 3 (p<0.01).

1789 A review of purity values and leukocyte fraction showed that the cluster is not characterized by a
1790 high level leukocytes or low level of purity.

1791

1792 Several interesting splicing events distinguished the adenocarcinoma-enriched cluster C2 and the
1793 squamous-enriched clusters C1 and C3 (Supplemental Table 29). *LIMK2* expression is associated
1794 with drug resistance in many tumor types and *LIMK2* knockdown has been shown to enhance
1795 chemotherapy effectiveness¹³⁵. The adenocarcinoma-enriched cluster showed stronger use of exon
1796 1 as the first exon (*LIMK2a* isoform) while the squamous clusters showed stronger use of exon 3
1797 as the first exon (*LIMK2b* isoform – missing the first LIM domain), suggesting alternate *LIMK2*
1798 expression regulation with potential impact on LIM mediated protein-protein interactions. Erbin
1799 is an adaptor protein produced by *ERBB2IP* that contributes to the oncogenic effects of HER2 and
1800 has also been a target of novel mutation specific immunotherapy^{136,137}. The *ERBB2IP* exon skip
1801 event removes the PDZ domain necessary for HER2 binding. Samples in C2 show PSI values
1802 40% lower than C1 and C3 samples, indicating that the Erbin expressed in C2 tumor samples is
1803 less capable of interacting with HER2. CD44 is a well-studied transmembrane glycoprotein with
1804 both oncogenic and tumor suppressor properties and splice variants that have been associated with
1805 metastatic progression^{138,139}. The adenocarcinoma-enriched C2 samples show reduced inclusion
1806 of the *CD44* variable exons 7-14 (generally referred to as v2-v9), which add extracellular stem
1807 structure with additional binding sites for posttranslational modifications and ligand-binding¹³⁸.

1808

1809 Less expected was the difference in splicing patterns that distinguish the C3 from C1 samples, as
1810 both predominantly contain squamous samples (Supplemental Table 30). The C3 cluster has a
1811 moderately better survival profile than the C1 cluster ($p < 0.05$; Supplemental Fig. S39). The splice
1812 events that most strongly distinguish C3 samples from C1 samples include several cancer related
1813 genes. *MAGI3* is a scaffold protein that regulates LPA to inhibit migration and invasion and
1814 cooperates with PTEN to modulate AKT kinase activity related to cell survival^{140,141}. The C3
1815 cluster includes the alternate exon 8 of *MAGI3* at a higher frequency (42% increase in PSI). Exon
1816 8 contains an annotated domain but codes for a 25-amino acid sequence between the second WW
1817 domain and the PDZ domain that interacts with PTEN, potentially altering protein interactions.
1818 *HACE1* is an E3 ubiquitin ligase that is a HER2 cooperative tumor suppressor¹⁴². Samples in C3
1819 include exon 7 at increased frequency (43% PSI increase). Exon 7 contains a premature stop codon
1820 leading to a truncated or degraded protein. Interestingly, many of the top splice events that define
1821 the C3 splicing cluster involve increased inclusion of an exon that introduces a premature stop
1822 codon or an alternate termination exon leading to a shortened protein product. These splicing
1823 events include *ABCC3_RI_16.2*, *MANBA_ES_2*, *DAPL1_AT_4*, *HACE1_ES_7*, and
1824 *TET2_RI_4.2*.

1825

1826

1827

1828

1829

1830 **S18. Batch Effect Analysis**

1831

1832 **Analysis Methods:**

1833 Hierarchical clustering and Principal Components Analysis (PCA) were used to assess batch
1834 effects in the CESC datasets. miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium
1835 HM450 microarray), mRNA sequencing (Illumina HiSeq), copy number variation (GW SNP 6),
1836 and protein expression (RPPA) datasets were analyzed across all CESC samples. All of the
1837 datasets were at TCGA level 3 since that is the level on which most of the analyses presented here
1838 are based. Batch effects were assessed with respect to two variables: batch ID and Tissue Source
1839 Site (TSS). Detailed results and batch effects analysis of other TCGA datasets can be found at
1840 <http://bioinformatics.mdanderson.org/tcgabatcheffects>.

1841

1842 For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson
1843 correlation coefficient as the dissimilarity measure. Samples were clustered and then annotated
1844 with colored bars at the bottom. Each color corresponds to a batch ID or a TSS. For PCA, we
1845 plotted the first four principal components, but only plots of the first two components are shown
1846 here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with
1847 centroids. Points representing samples with the same batch ID (or TSS) were connected to the
1848 batch centroid by lines. The centroids were computed by taking the mean across all samples in
1849 the batch. That procedure produced a visual representation of the relationships among batch
1850 centroids in relation to the scatter within batches.

1851 **miRNA Results**

1852 Supplemental Figure S40 shows clustering and PCA plots for miRNA-seq data. miRNAs with
1853 zero values were removed and the read counts were \log_2 -transformed before generating the figures.
1854 The figures show small batch effects by both batch ID and TSS; however, the magnitude of batch
1855 effects wasn't high and we did not believe that it warranted batch effects correction and subsequent
1856 potential loss of important biological and technical variation in the data.

1857

1858 **DNA Methylation Results**

1859 Supplemental Figure S41 shows clustering and PCA plots for the Infinium DNA methylation
1860 platform. Small batch effects by batch ID and TSS were seen, but once again they were
1861 deemed small enough not to warrant batch effects correction.

1862

1863 **RNA-seqV2 Results**

1864 Supplemental Figure S42 shows clustering and PCA plots for the RNA-seq platform. Small batch
1865 effects were seen by both batch ID and TSS, but not enough to warrant algorithmic batch effects
1866 correction.

1867

1868

1869

1870 **Copy Number Variation Results**

1871 Supplemental Figure S43 shows clustering and PCA plots for the copy number variation data
1872 generated on the SNP 6 platform. Small batch effects were seen by both batch ID and TSS, but
1873 not enough to warrant algorithmic batch effects correction.

1874

1875 **Protein Expression Results**

1876 Supplemental Figure S44 shows clustering and PCA plots for the protein expression data generated
1877 on the RPPA platform. Small batch effects were seen by both batch ID and TSS, but not enough
1878 to warrant algorithmic batch effects correction.

1879

1880

1881

1882

1883

1884

1885

1886

1887 **References**

- 1888 68. Cancer, I.A.f.R.o. WHO Classification of Tumours of Female Reproductive Organs Vol.
1889 6 (International Agency for Research on Cancer 2014)
- 1890 69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
1891 transform. *Bioinformatics* **25**, 1754-1760 (2009)
- 1892 70. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
1893 database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997)
- 1894 71. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**,
1895 2078-2079 (2009)
- 1896 72. Burk, R.D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology*
1897 **445**, 232-243 (2013)
- 1898 73. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
1899 thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006)
- 1900 74. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation
1901 sequencing of tumors. *Bioinformatics* **26**, 730-736 (2010)
- 1902 75. Simpson, J.T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome*
1903 *Res.* **19**, 1117-1123 (2009)
- 1904 76. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**,
1905 909-912 (2010)
- 1906 77. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656-664 (2002)
- 1907 78. Kuhn, R.M., Haussler, D. & Kent, W.J. The UCSC genome browser and associated
1908 tools. *Brief. Bioinform.* **14**, 144-161 (2013)
- 1909 79. Schwartz, S. Papillomavirus transcripts and posttranscriptional regulation. *Virology* **445**,
1910 187-196 (2013)
- 1911 80. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of
1912 insertions, deletions and gene fusions. *Genome Biol.* **14**, R36-R36 (2013)
- 1913 81. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth
1914 approach to detect break points of large deletions and medium sized insertions from
1915 paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009)
- 1916 82. Johansson, C. & Schwartz, S. Regulation of human papillomavirus gene expression by
1917 splicing and polyadenylation. *Nat. Rev. Microbiol.* **11**, 239-251 (2013)
- 1918 83. Radenbaugh, A.J. *et al.* RADIA: RNA and DNA Integrated Analysis for somatic
1919 mutation detection. *PLoS One* **9**, e111516 (2014)
- 1920 84. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types.
1921 *Nature* **502**, 333-339 (2013)
- 1922 85. Dabney, A.R. ClaNC: point-and-click software for classifying microarrays to nearest
1923 centroids. *Bioinformatics* **22**, 122-123 (2006)
- 1924 86. de Hoon, M.J.L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software.
1925 *Bioinformatics* **20**, 1453-1454 (2004)

- 1926 87. Saldanha, A.J. Java Treeview—extensible visualization of microarray data.
1927 *Bioinformatics* **20**, 3246-3248 (2004)
- 1928 88. Kim, D. & Salzberg, S.L. TopHat-Fusion: an algorithm for discovery of novel fusion
1929 transcripts. *Genome Biol.* **12**, R72-R72 (2011)
- 1930 89. Chen, K. *et al.* BreakFusion: targeted assembly-based identification of gene fusions in
1931 whole transcriptome paired-end sequencing data. *Bioinformatics* **28**, 1923-1924 (2012)
- 1932 90. Greger, L. *et al.* Tandem RNA chimeras contribute to transcriptome diversity in human
1933 population and are associated with intronic genetic variants. *PLoS One* **9**, e104567 (2014)
- 1934 91. Torres-García, W. *et al.* PRADA: pipeline for RNA sequencing data analysis.
1935 *Bioinformatics* **30**, 2224-2226 (2014)
- 1936 92. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated
1937 transcript fusions. *Oncogene* (2014)
- 1938 93. Chen, K. *et al.* TIGRA: A targeted iterative graph routing assembler for breakpoint
1939 assembly. *Genome Res.* **24**, 310-317 (2014)
- 1940 94. Parker, B.C. *et al.* The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a
1941 regulation in glioblastoma. *J. Clin. Invest.* **123**, 855-865 (2013)
- 1942 95. Godinho, M., Meijer, D., Setyono-Han, B., Dorssers, L.C.J. & van Agthoven, T.
1943 Characterization of BCAR4, a novel oncogene causing endocrine resistance in human
1944 breast cancer cells. *J. Cell. Physiol.* **226**, 1741-1749 (2011)
- 1945 96. Xing, Z. *et al.* lncRNA directs cooperative epigenetic regulation downstream of
1946 chemokine signals. *Cell* **159**, 1110-1125 (2014)
- 1947 97. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology.
1948 *Epigenomics* **3**, 771-784 (2011)
- 1949 98. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of
1950 lung adenocarcinoma. *Nature* **511**, 543-550 (2014)
- 1951 99. The Cancer Genome Atlas Network. Comprehensive molecular characterization of
1952 human colon and rectal cancer. *Nature* **487**, 330-337 (2012)
- 1953 100. Smyth, G.K. *Limma: linear models for microarray data.* (Springer, New York, New
1954 York, USA; 2005)
- 1955 101. Wilks, C. *et al.* The Cancer Genomics Hub (CGHub): overcoming cancer through the
1956 power of torrential data. *Database* **2014** (2014)
- 1957 102. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization.
1958 *BMC Bioinformatics* **11**, 367-367 (2010)
- 1959 103. Li, J. & Tibshirani, R. Finding consistent patterns: A nonparametric approach for
1960 identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22**, 519-
1961 536 (2013)
- 1962 104. Mullokandov, G. *et al.* High-throughput assessment of microRNA activity and function
1963 using microRNA sensor and decoy libraries. *Nat. Methods* **9**, 840-846 (2012)
- 1964 105. Tay, Y., Rinn, J. & Pandolfi, P.P. The multilayered complexity of ceRNA crosstalk and
1965 competition. *Nature* **505**, 344-352 (2014)
- 1966 106. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
1967 *Bioinformatics* **28**, 1353-1358 (2012)

- 1968 107. Xie, H. *et al.* Novel functions and targets of miR-944 in human cervical cancer cells. *Int. J. Cancer* **136**, E230-E241 (2015)
- 1969
- 1970 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)
- 1971
- 1972
- 1973 109. Hu, J. *et al.* Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986-1994 (2007)
- 1974
- 1975 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)
- 1976
- 1977 111. Ju, Z. *et al.* Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics* **31**, 912-918 (2015)
- 1978
- 1979 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)
- 1980
- 1981
- 1982 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)
- 1983
- 1984
- 1985 114. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010)
- 1986
- 1987 115. Moris, A., Murray, S.M. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Front. Microbiol.* **5** (2014)
- 1988
- 1989 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)
- 1990
- 1991
- 1992 117. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685-D690 (2011)
- 1993
- 1994 118. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **74**, 016110 (2006)
- 1995
- 1996 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)
- 1997
- 1998
- 1999 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)
- 2000
- 2001
- 2002 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)
- 2003
- 2004
- 2005 122. Patel, S.P. & Kurzrock, R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Mol. Cancer Ther.* **14**, 847-856 (2015)
- 2006
- 2007 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)
- 2008
- 2009

2010 124. Brahmer, J.R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with
2011 advanced cancer. *New Engl. J. Med.* **366**, 2455-2465 (2012)

2012 125. Topalian, S.L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in
2013 cancer. *New Engl. J. Med.* **366**, 2443-2454 (2012)

2014 126. Dihlmann, S. *et al.* Lack of Absent in Melanoma 2 (AIM2) expression in tumor cells is
2015 closely associated with poor survival in colorectal cancer patients. *Int. J. Cancer* **135**,
2016 2387-2396 (2014)

2017 127. de Koning, H.D., van Vlijmen-Willems, I.M., Zeeuwen, P.L., Blokkx, W.A & Schalkwijk,
2018 J. Absent in Melanoma 2 is predominantly present in primary melanoma and primary
2019 squamous cell carcinoma, but largely absent in metastases of both tumors. *J. Am. Acad.*
2020 *Dermatol.* **71**, 1012-1015 (2014)

2021 128. Unterholzner, L. *et al.* IFI16 is an innate immune sensor for intracellular DNA. *Nat.*
2022 *Immunol.* **11**, 997-1004 (2010)

2023 129. Reinholz, M. *et al.* HPV16 activates the AIM2 inflammasome in keratinocytes. *Arch.*
2024 *Dermatol. Res.* **305**, 723-732 (2013)

2025 130. Strieter, R.M. *et al.* Cancer CXCR chemokine networks and tumour angiogenesis. *Eur. J.*
2026 *Cancer* **42**, 768-778 (2006)

2027 131. Naidoo, J., Page, D.B. & Wolchok, J.D. Immune modulation for cancer therapy. *Br. J.*
2028 *Cancer* **111**, 2214-2219 (2014)

2029 132. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev.*
2030 *Cancer* **12**, 252-264 (2012)

2031 133. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-
2032 generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011)

2033 134. Ryan, M.C., Cleland, J., Kim, R., Wong, W.C. & Weinstein, J.N. SpliceSeq: a resource
2034 for analysis and visualization of RNA-Seq data on alternative splicing and its functional
2035 impacts. *Bioinformatics* **28**, 2385-2387 (2012)

2036 135. Gamell, C., Schofield, A.V., Suryadinata, R., Sarcevic, B. & Bernard, O. LIMK2
2037 mediates resistance to chemotherapeutic drugs in neuroblastoma cells through regulation
2038 of drug-induced cell cycle arrest. *PLoS One* **8**, e72850 (2013)

2039 136. Tao, Y. *et al.* Role of Erbin in ErbB2-dependent breast tumor growth. *Proc. Natl. Acad.*
2040 *Sci. USA* **111**, E4429-E4438 (2014)

2041 137. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a
2042 patient with epithelial cancer. *Science* **344**, 641-645 (2014)

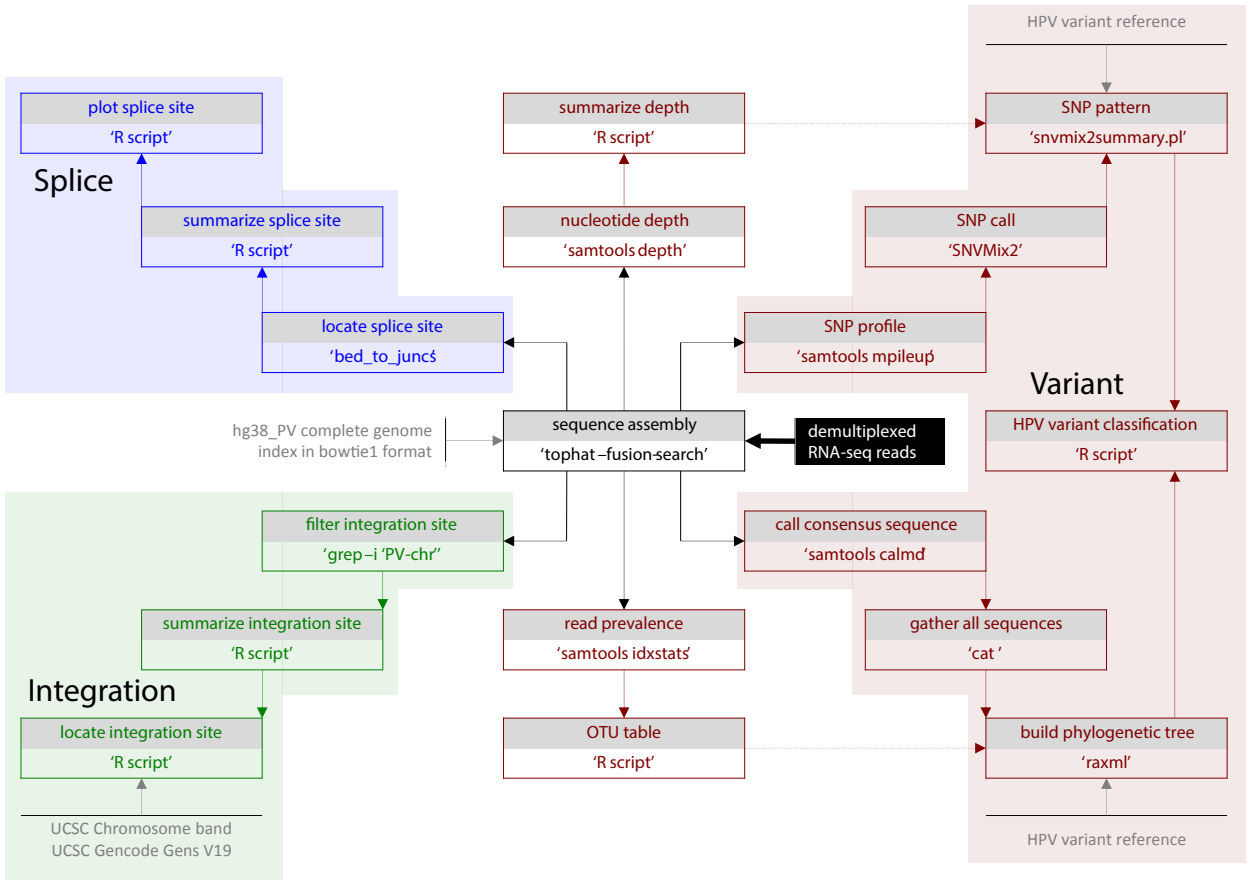
2043 138. Louderbough, J.M.V. & Schroeder, J.A. Understanding the dual nature of CD44 in
2044 breast cancer progression. *Mol. Cancer Res.* **9**, 1573-1586 (2011)

2045 139. Speiser, P. *et al.* CD44 is an independent prognostic factor in early-stage cervical cancer.
2046 *Int. J. Cancer* **74**, 185-188 (1997)

2047 140. Lee, S.J. *et al.* MAGI-3 competes with NHERF-2 to negatively regulate LPA2 receptor
2048 signaling in colon cancer cells. *Gastroenterology* **140**, 924-934 (2011)

2049 141. Wu, Y. *et al.* Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of
2050 MAGI3, a novel membrane-associated guanylate kinase. *J. Biol. Chem.* **275**, 21477-
2051 21485 (2000)

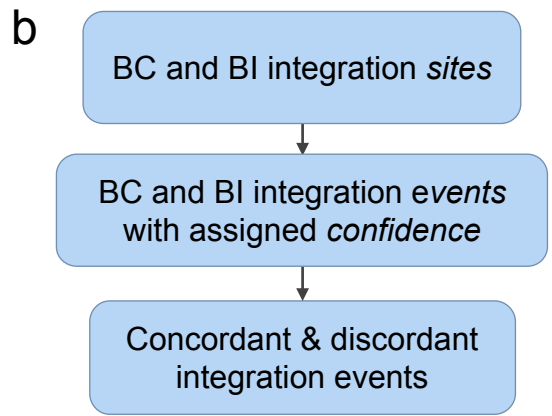
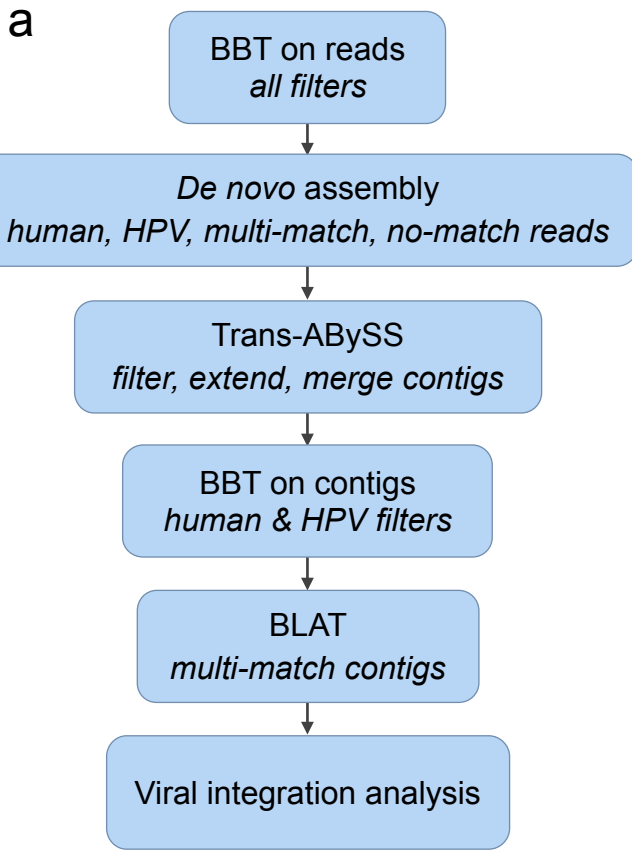
2052 142. Goka, E.T. & Lippman, M.E. Loss of the E3 ubiquitin ligase HACE1 results in enhanced
2053 Rac1 signaling contributing to breast cancer progression. *Oncogene* (2015)
2054
2055



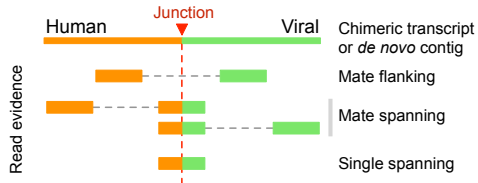
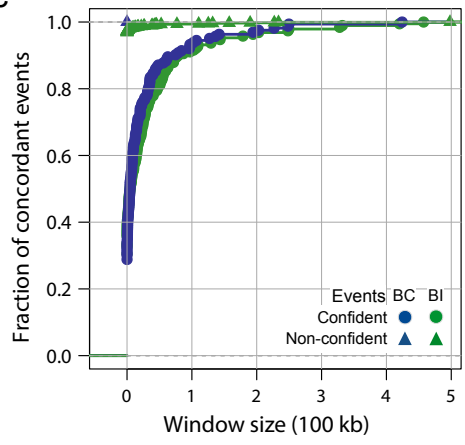
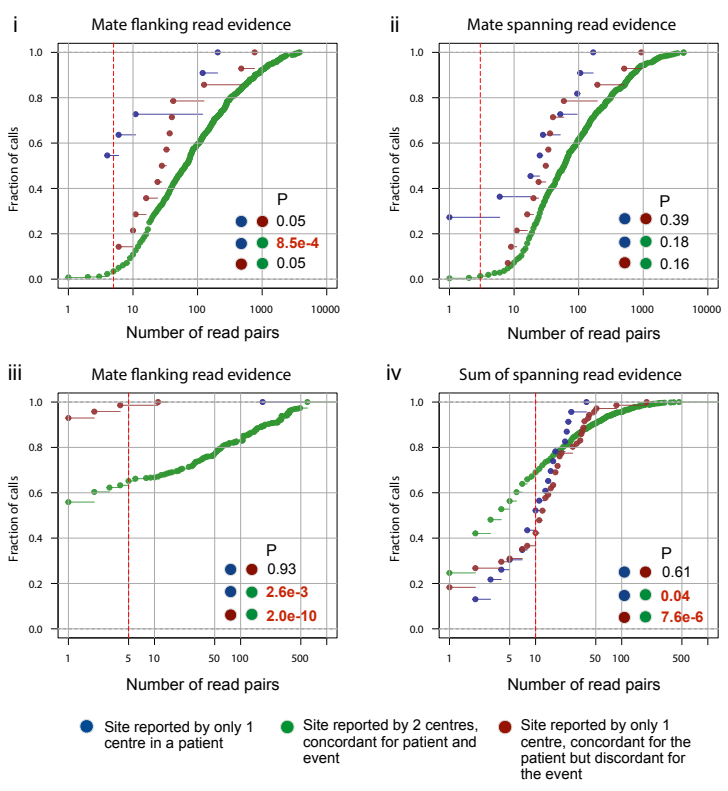
Supplemental Fig. S1

Supplemental Figure S1: HPV16 variant calling analysis pipeline.

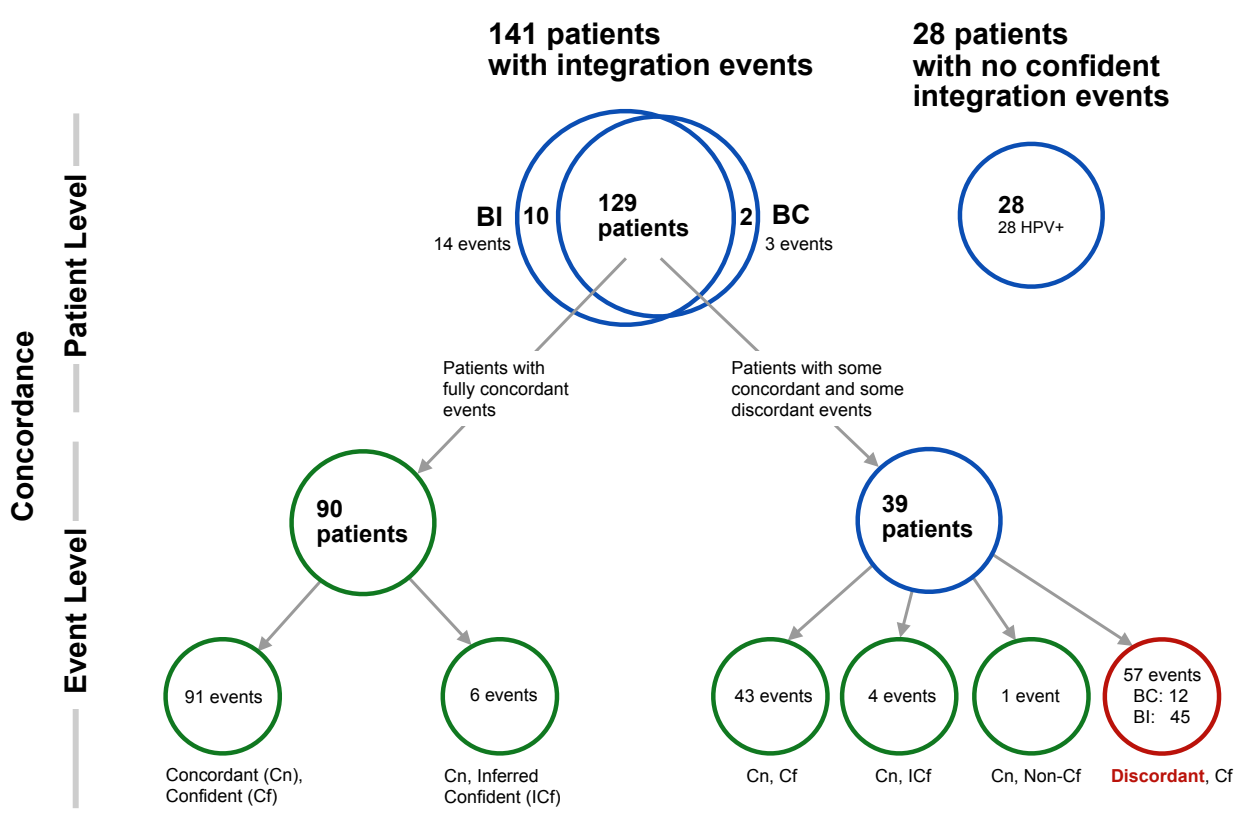
Supplemental Figure S2: Examples of HPV16 reads that indicate unspliced (a) and spliced (b) E6 transcripts.



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

a**c****b**

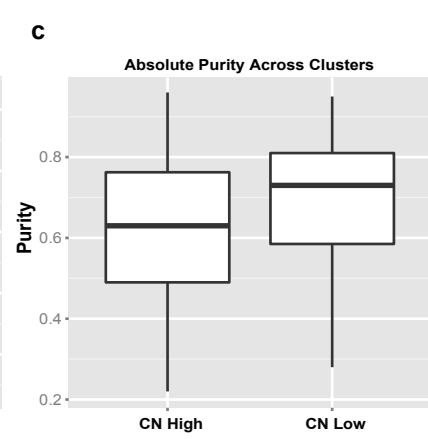
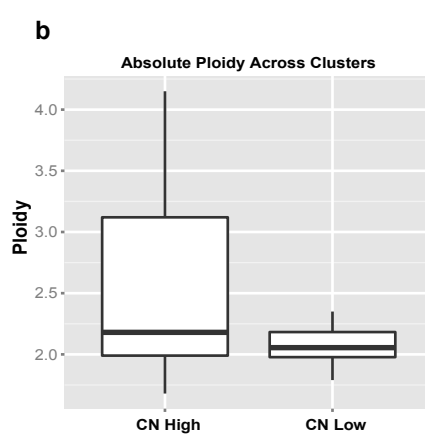
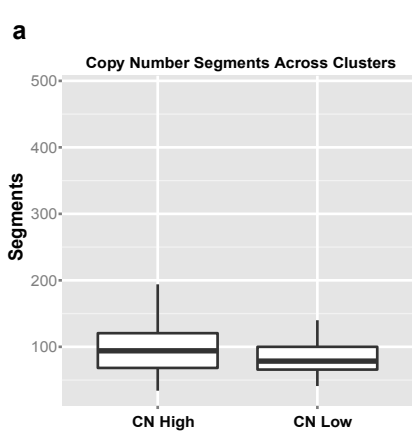
Supplemental Figure S4: Concordance analysis for integration events for RNAseq data. **a**, Types of read evidence for integration sites from RNAseq data. **b**, Distribution functions for flanking and spanning read evidence, with dashed red lines showing evidence thresholds. BC: **i**) mate flanking = 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.



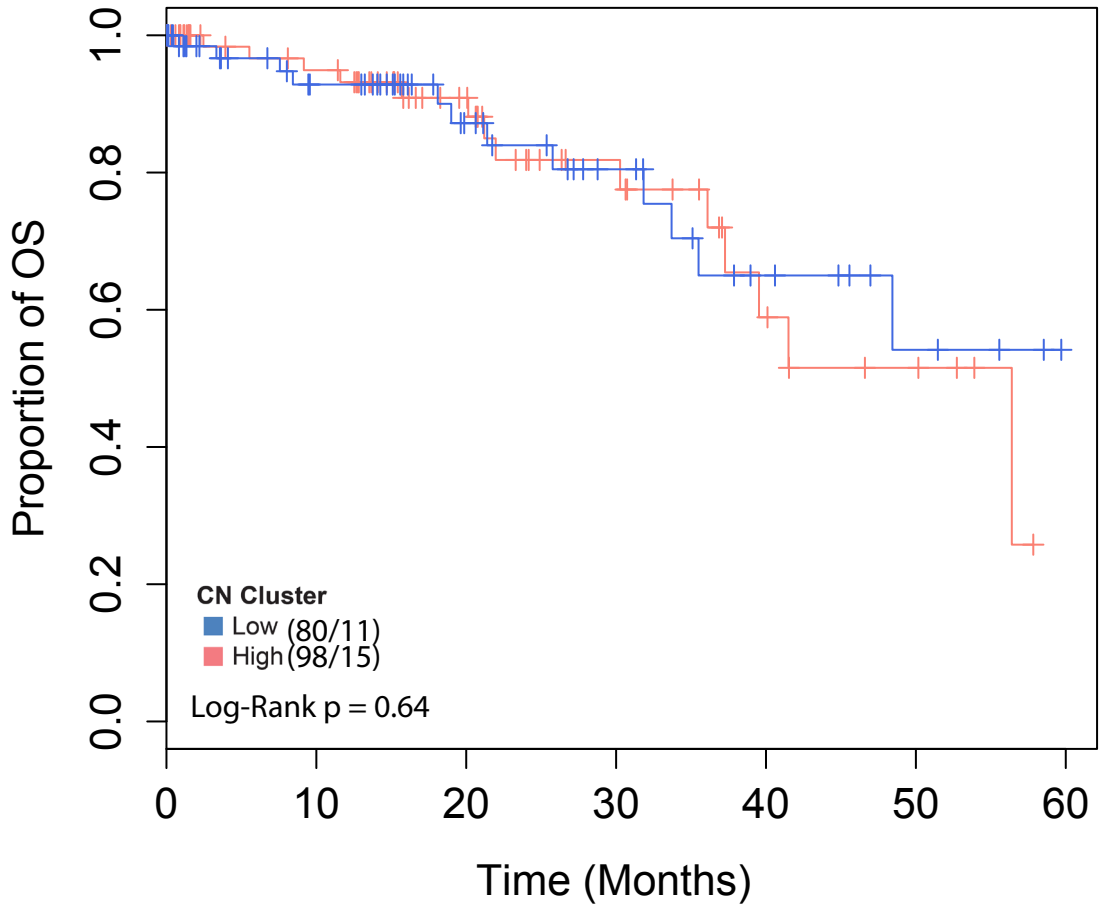
129 of 141 (91%) patients, 145 Concordant integration events

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.

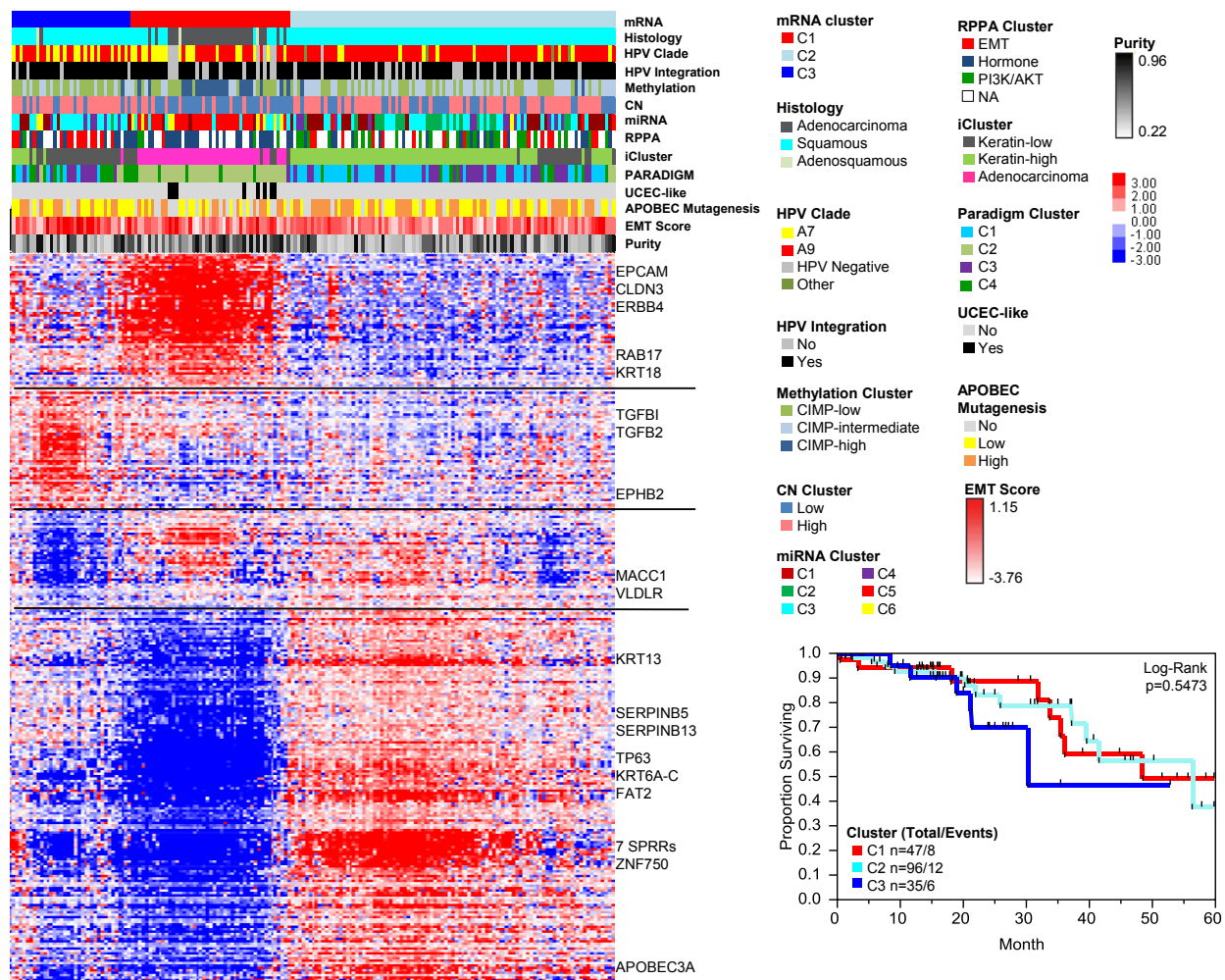
Supplemental Figure S6: Significantly Mutated Genes. a-g, High-confidence somatic mutations in significantly mutated genes (SMGs) among 192 exome-sequenced samples in the Extended dataset are shown. SMGs presented in Extended Data Fig. 2 are not shown here. Domains are labeled in accordance with Gencode 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.

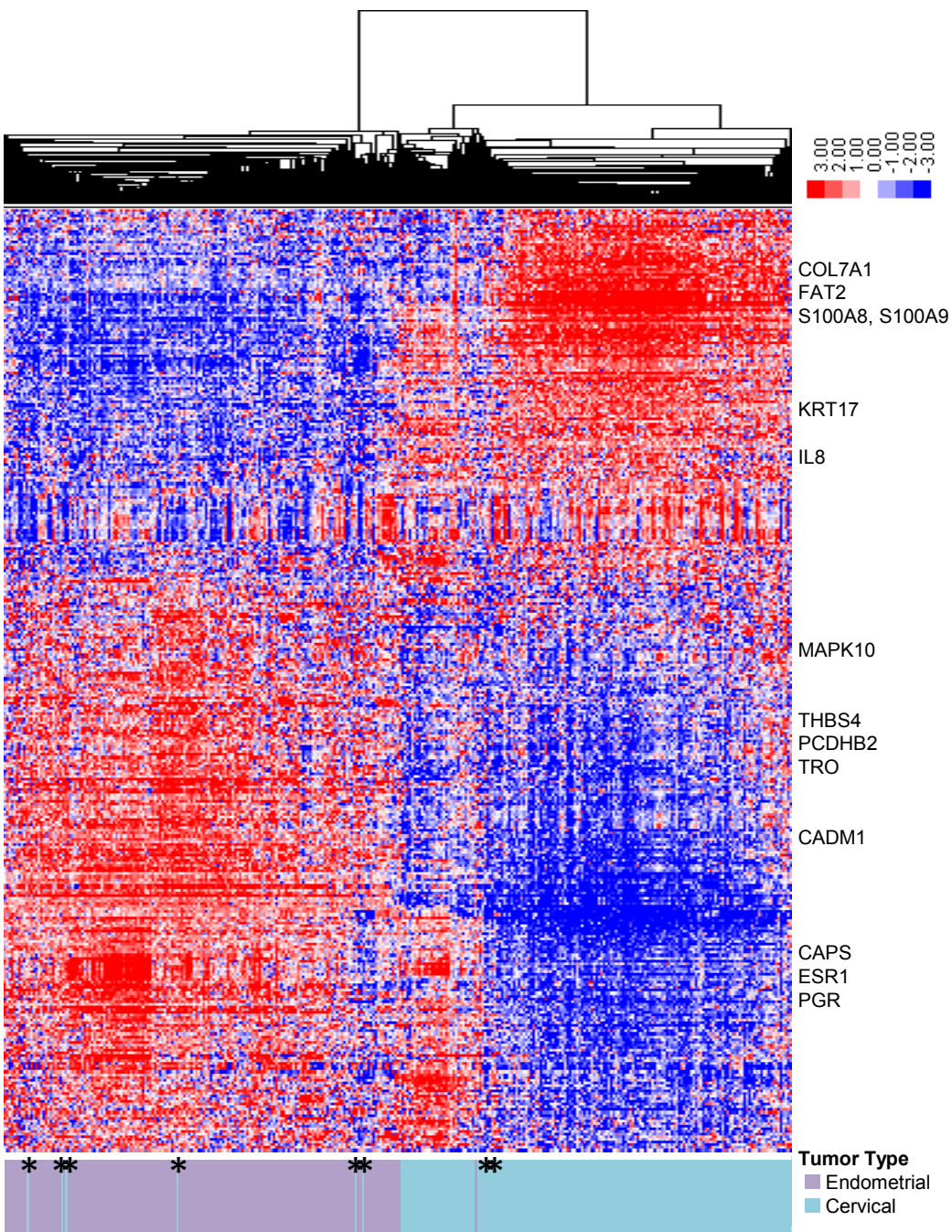


Supplemental Figure S8: Survival analysis between CN clusters. Kaplan-Meier survival analysis between cases within the CN High (High) and CN Low (Low) clusters.



Supplemental Fig. S9

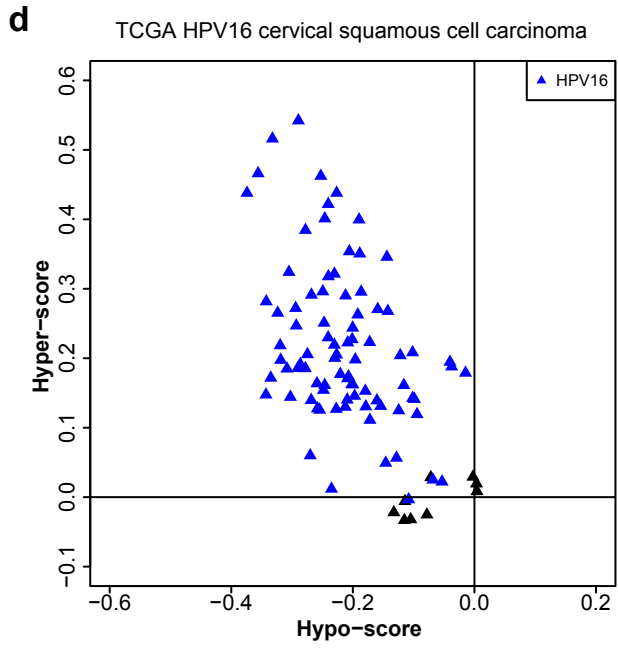
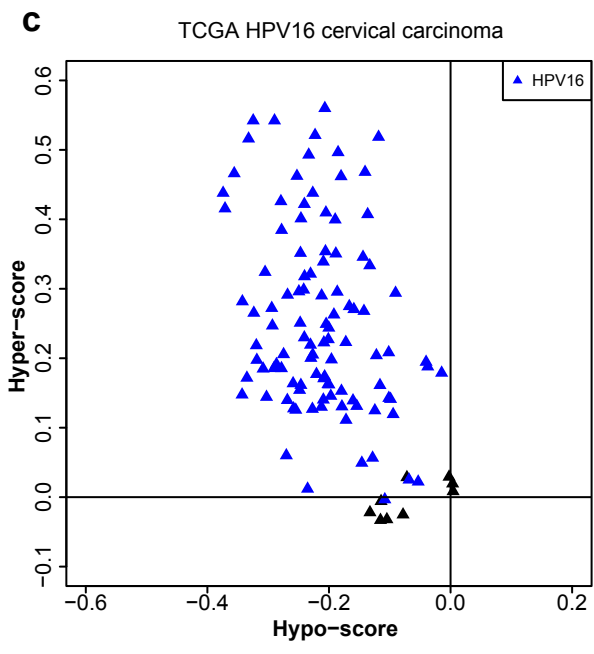
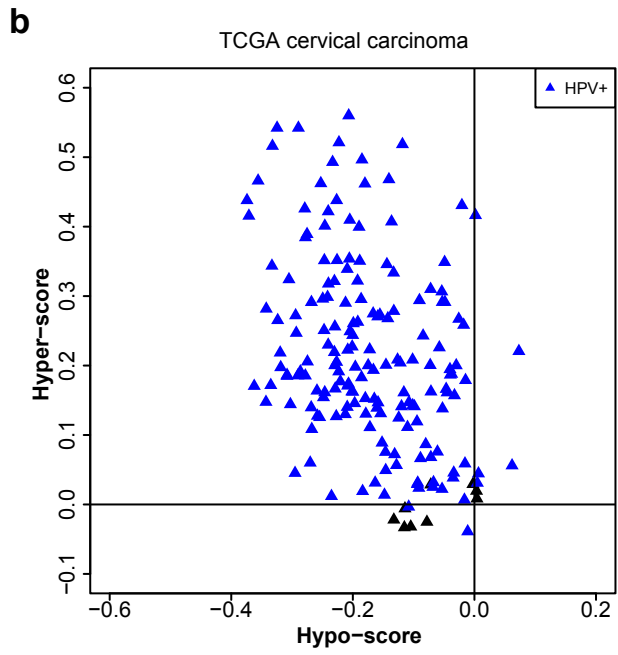
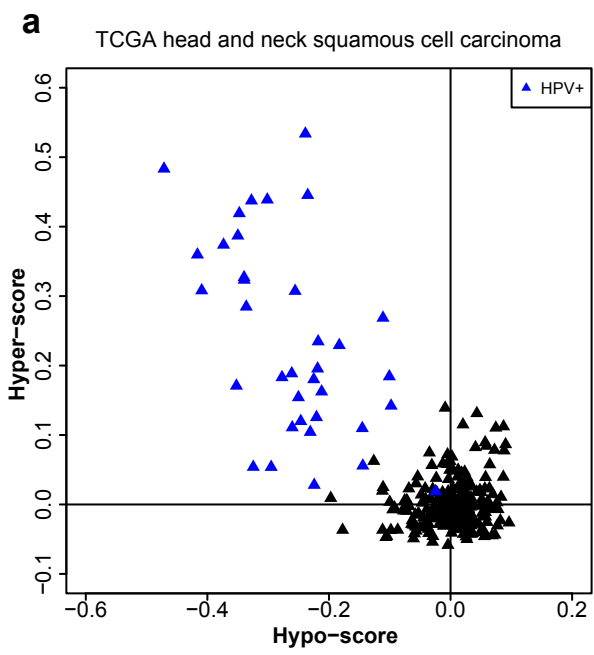
Supplemental Figure S9: mRNA clustering analysis. Gene expression values obtained from RNAseq data on 300 signature genes (y-axis) across 178 cervical cancer samples (x-axis) were hierarchically clustered using uncentered correlation and centroid linkage as the clustering method (left). Normalized gene 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. Horizontal black lines approximately separate gene clusters. Five-year survival analysis of cervical cancer patients grouped according to mRNA cluster membership (lower right panel).



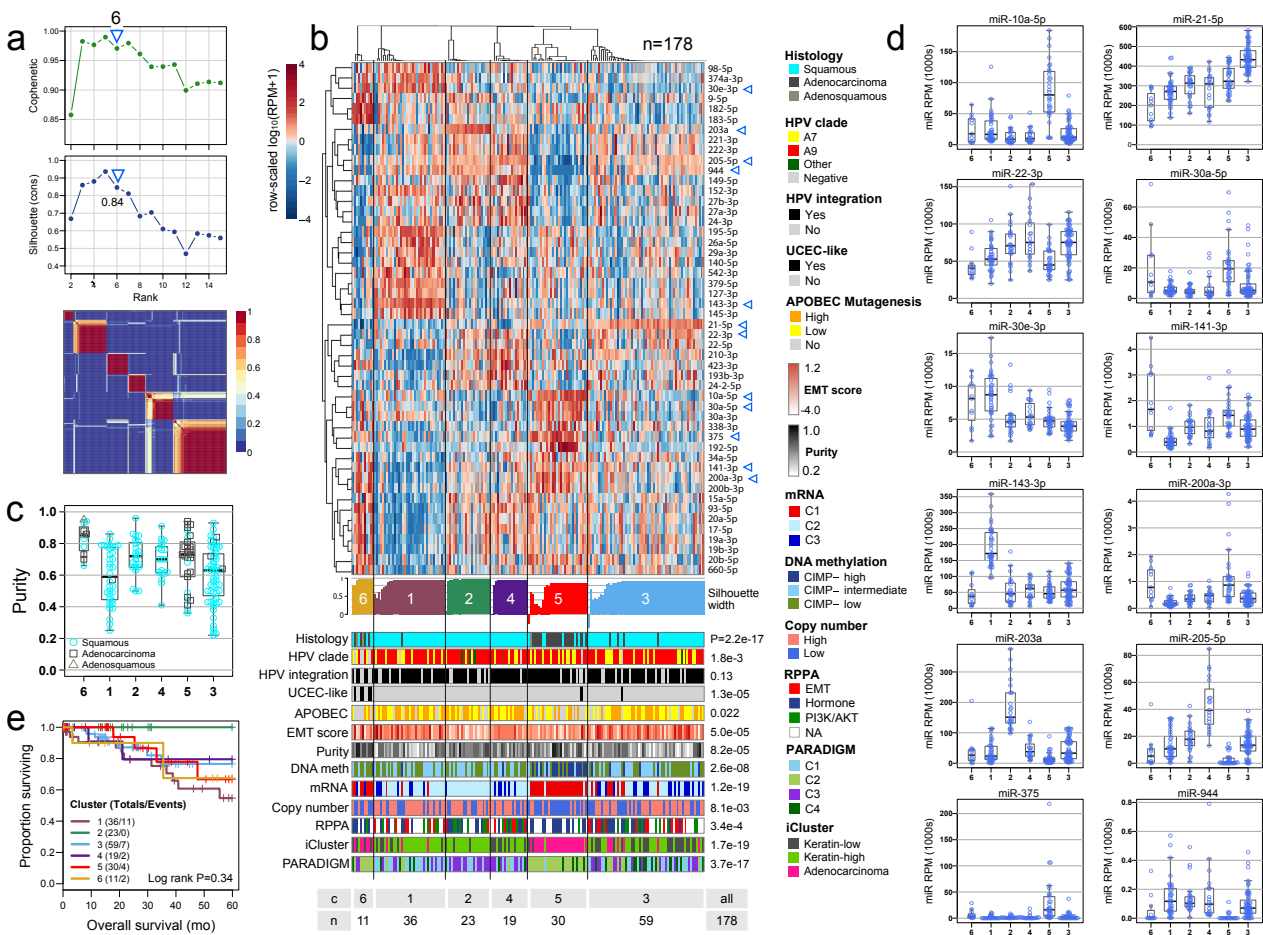
*: UCEC-like samples

Supplemental Figure S10: Clustering analysis of CESC and UCEC TCGA samples.

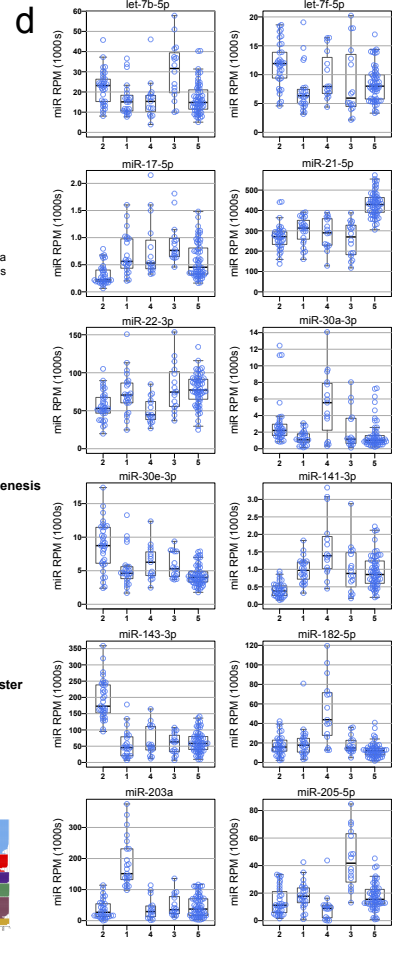
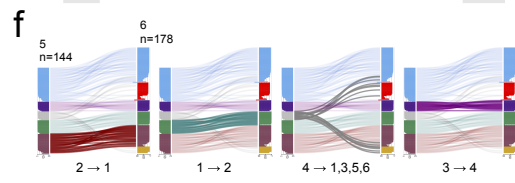
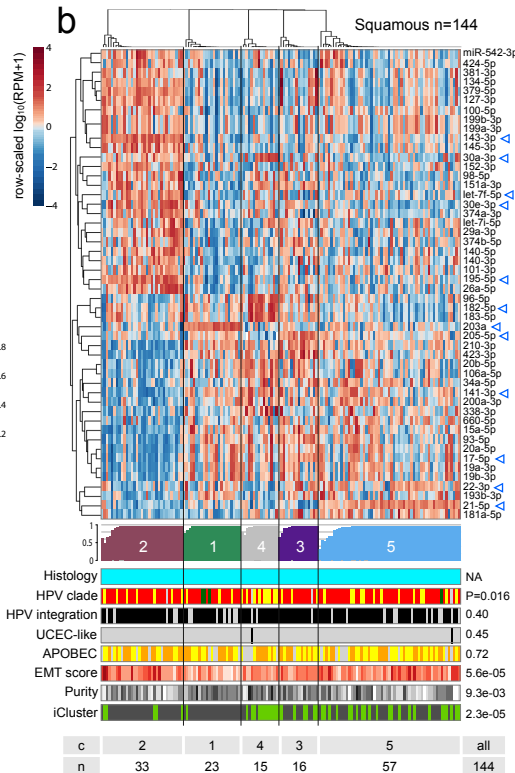
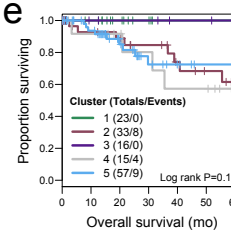
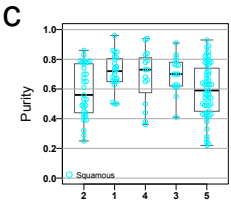
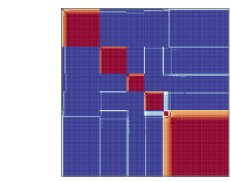
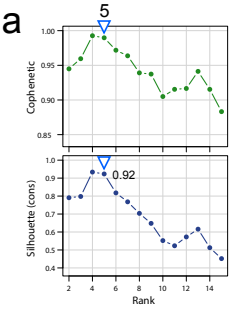
ANOVA was performed on normalized gene-level RSEM values for 178 cervical cancer and 170 TCGA endometrial cancer samples to identify differentially expressed genes between the two cancer types. The differentially expressed genes (n=384, FDR <0.05) and samples were clustered using uncentered correlation 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.

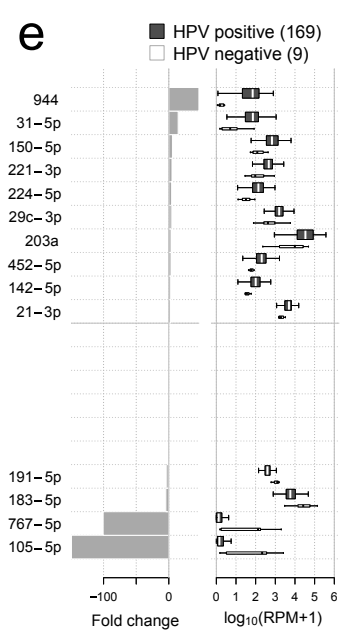
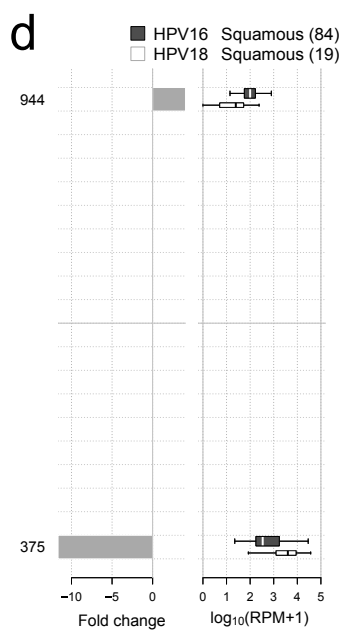
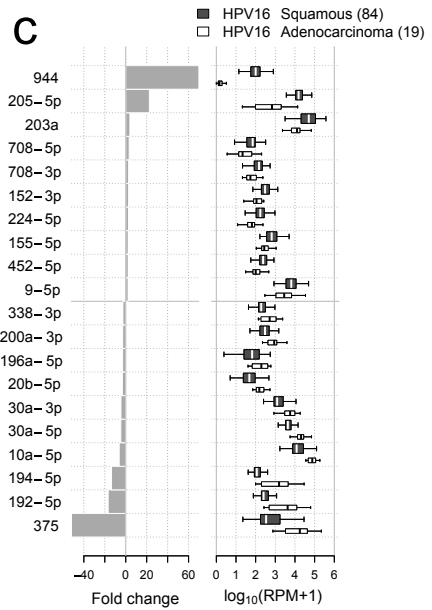
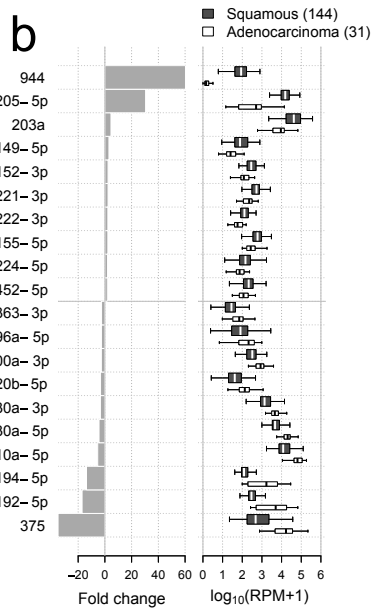
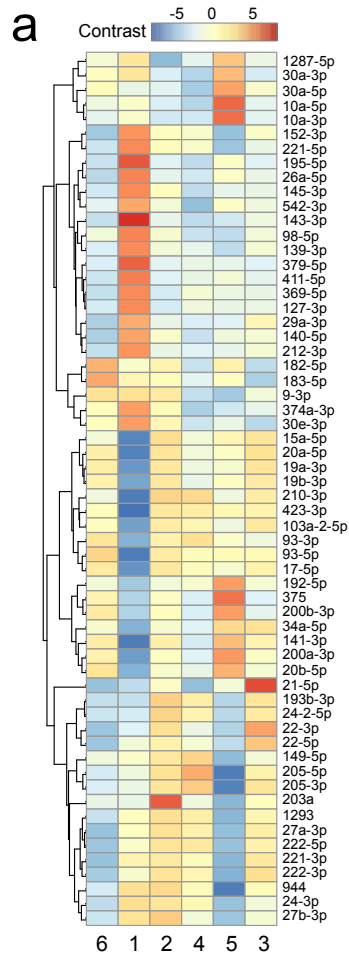


Supplemental Figure S12: Unsupervised NMF consensus clustering of miRNA mature strand data. **a**, Rank survey profiles for cophenetic correlation coefficient and average silhouette width for the NMF consensus clustering rank survey for 178 tumor samples, and a blue/red heatmap showing sample consensus 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 $\log_{10}(\text{RPM}+1)$ 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.

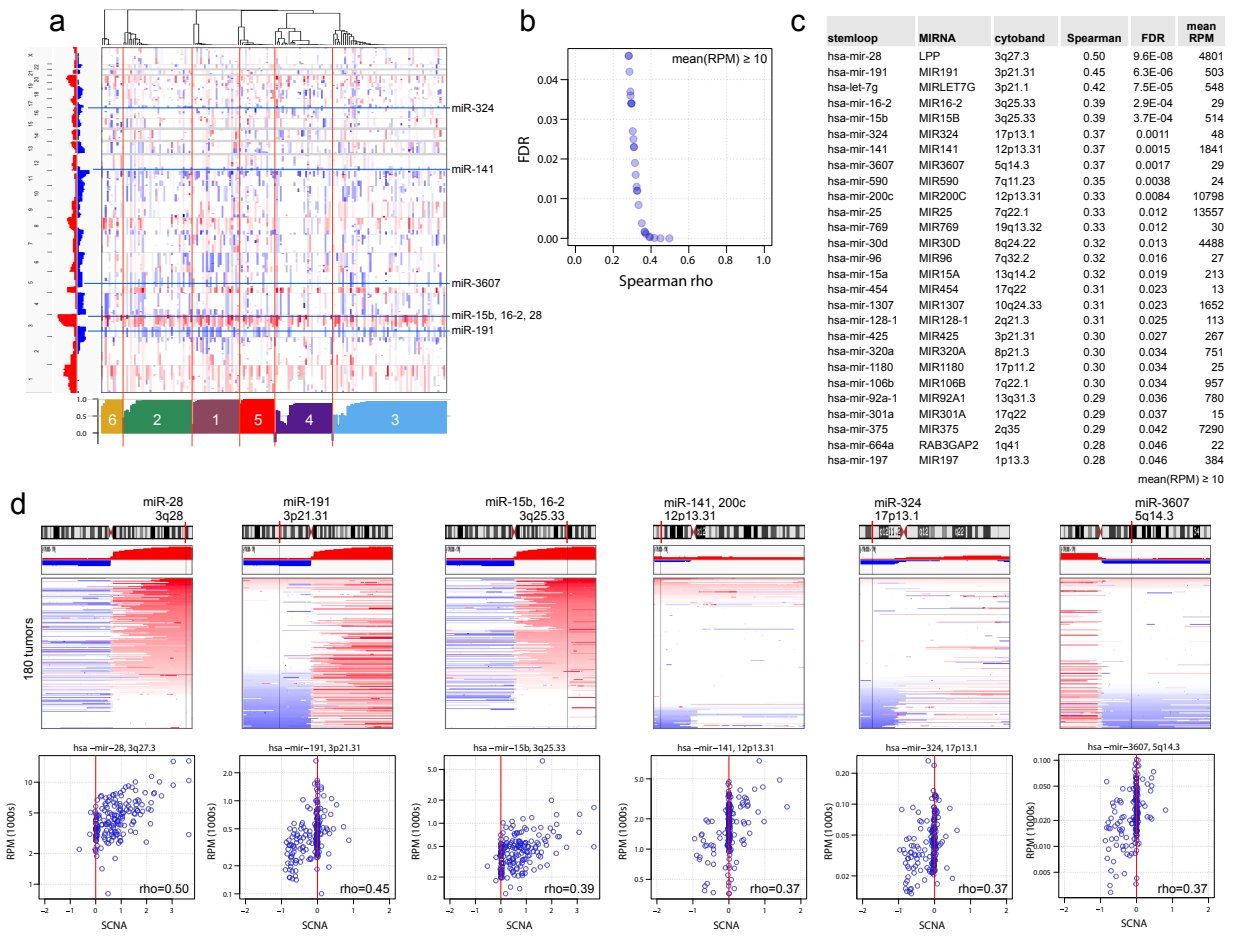


Supplemental Fig. S13

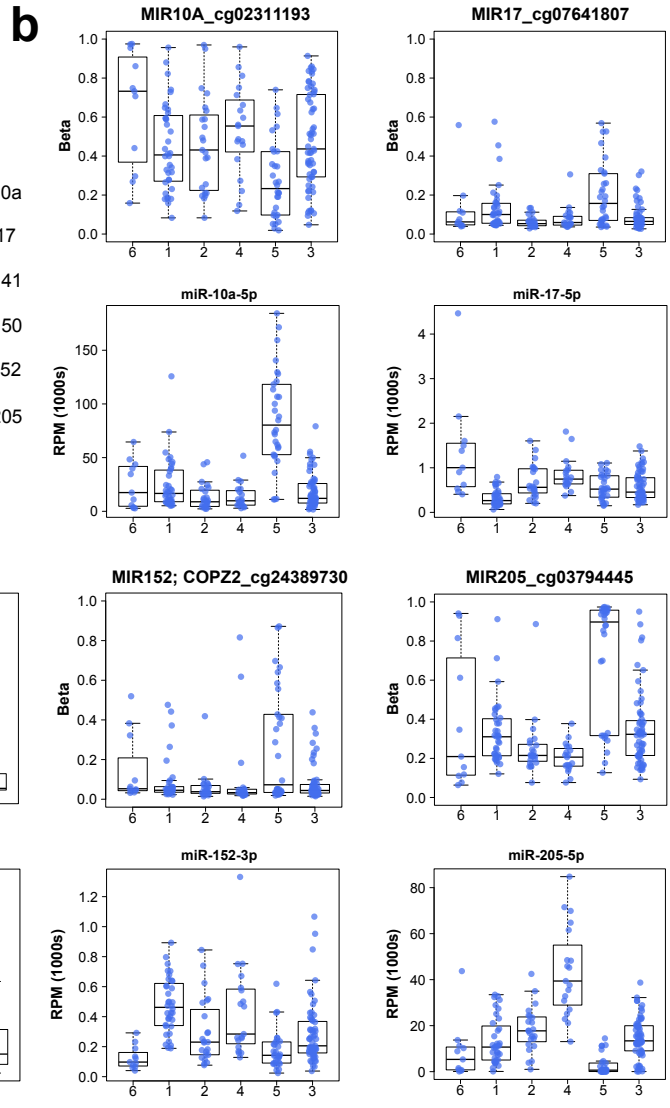
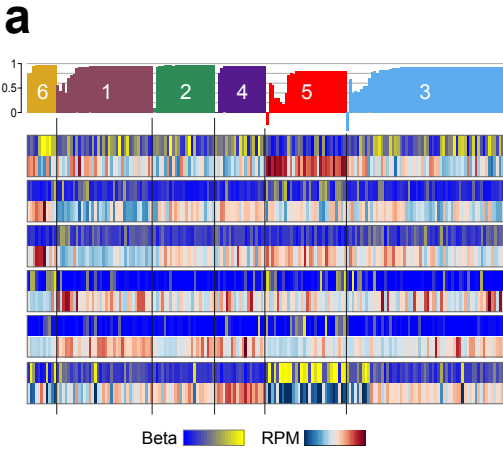
Supplemental Figure S13: Unsupervised NMF consensus clustering of miRNA mature strand data for squamous tumors (n=144). **a-e**, Panels are as in Supplemental Figure S12. **f**, Relationships between sample locations across the current 5-cluster solution for n=144, and the 6-cluster solution for n=178 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**.



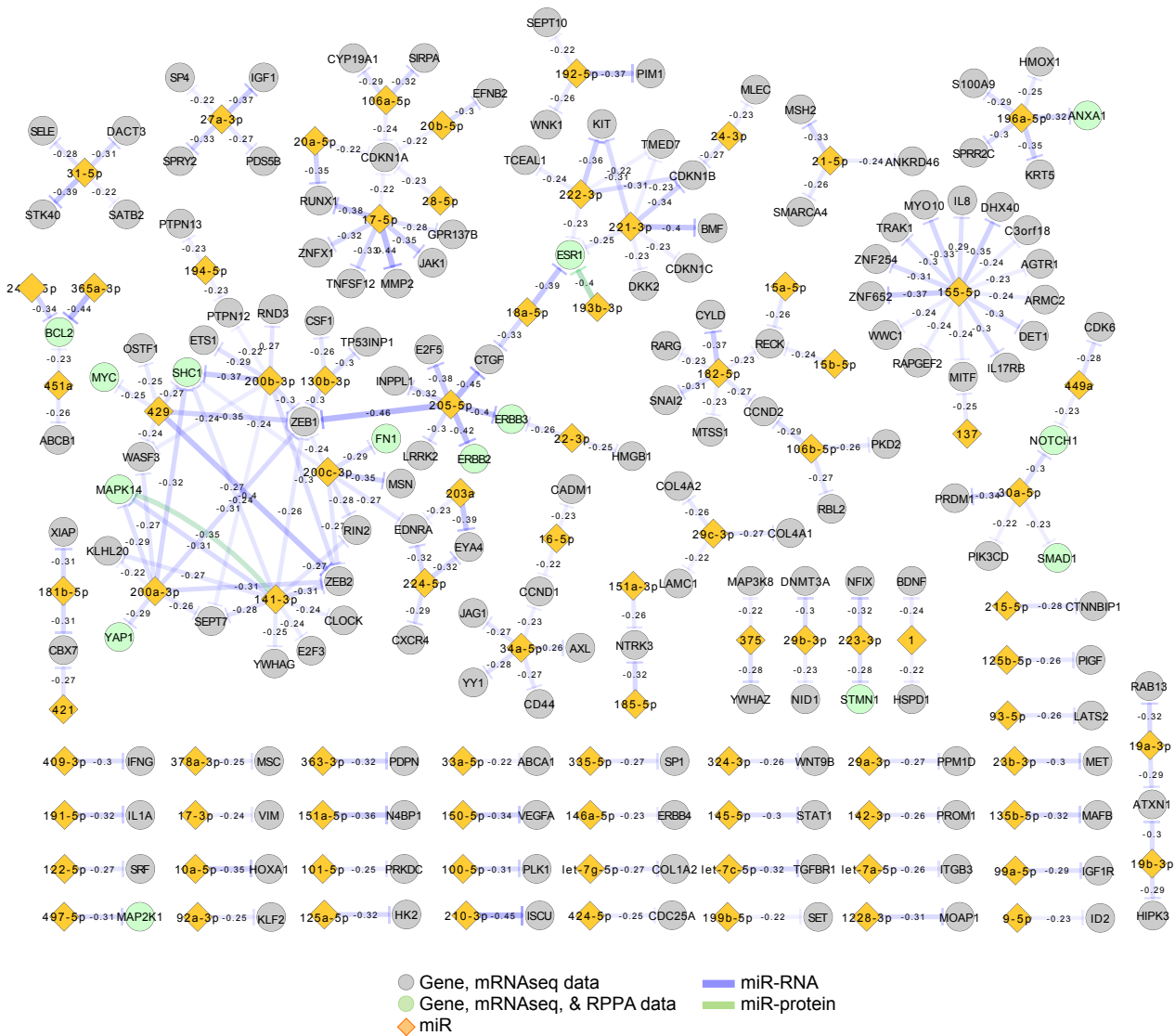
Supplemental Figure S14: Differentially abundant miRs. **a**, Heatmap of differential abundance contrasts for the 50 miRs that were scored highly by a SAMseq multiclass analysis across the six unsupervised clusters. **b-e**, miRs with the largest fold-changes for b) squamous vs. adenocarcinomas, c) HPV16 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.



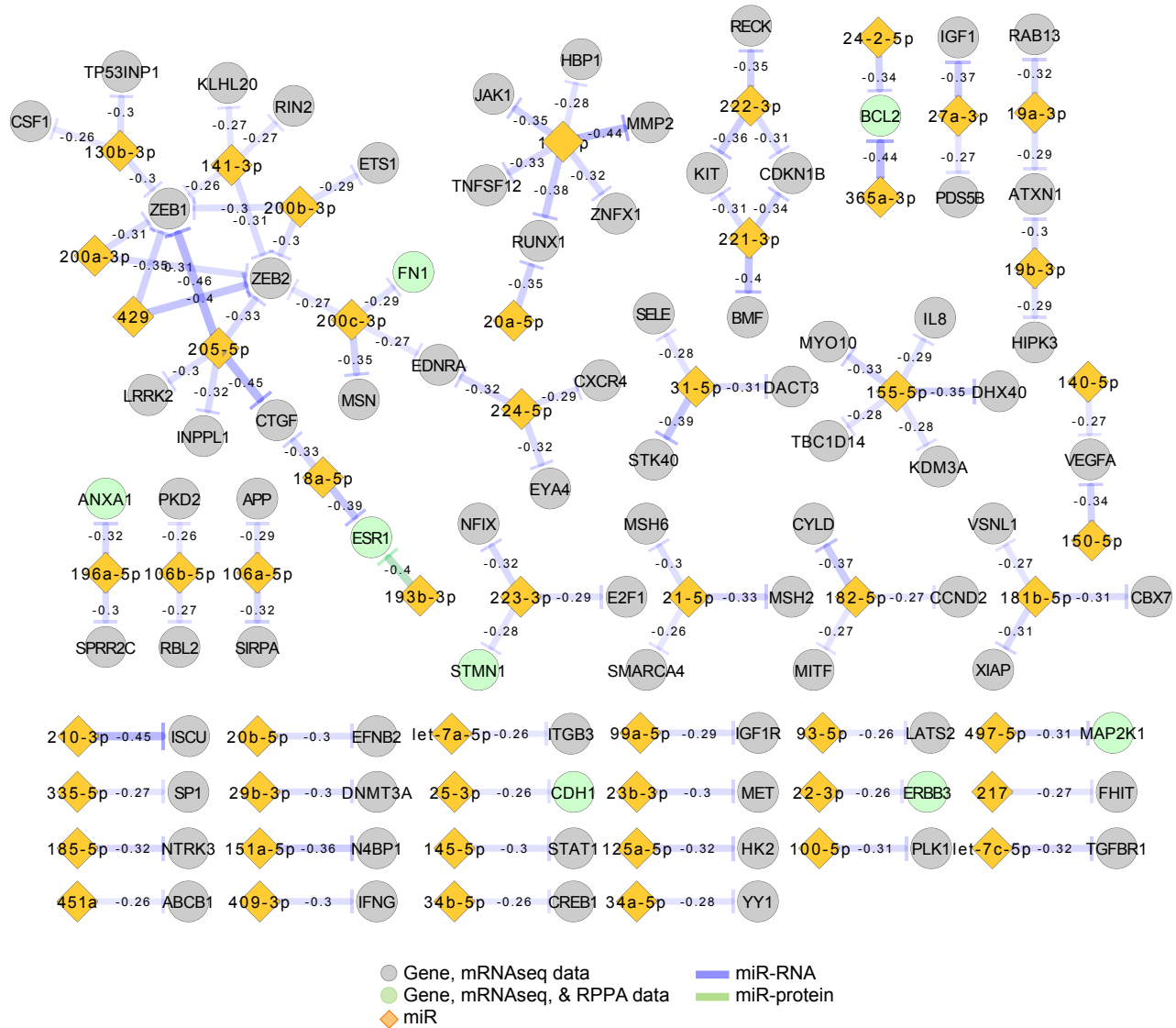
Supplemental Figure S15: Relationships between somatic copy number (SCNA) and pre-miRNA abundance. **a**, Global relationship between SCNA and miR-based unsupervised clusters. Blue horizontal lines mark the locations of the miRNAs in panel **d**. **b**, Relationship between the Spearman correlation coefficient (ρ) 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.



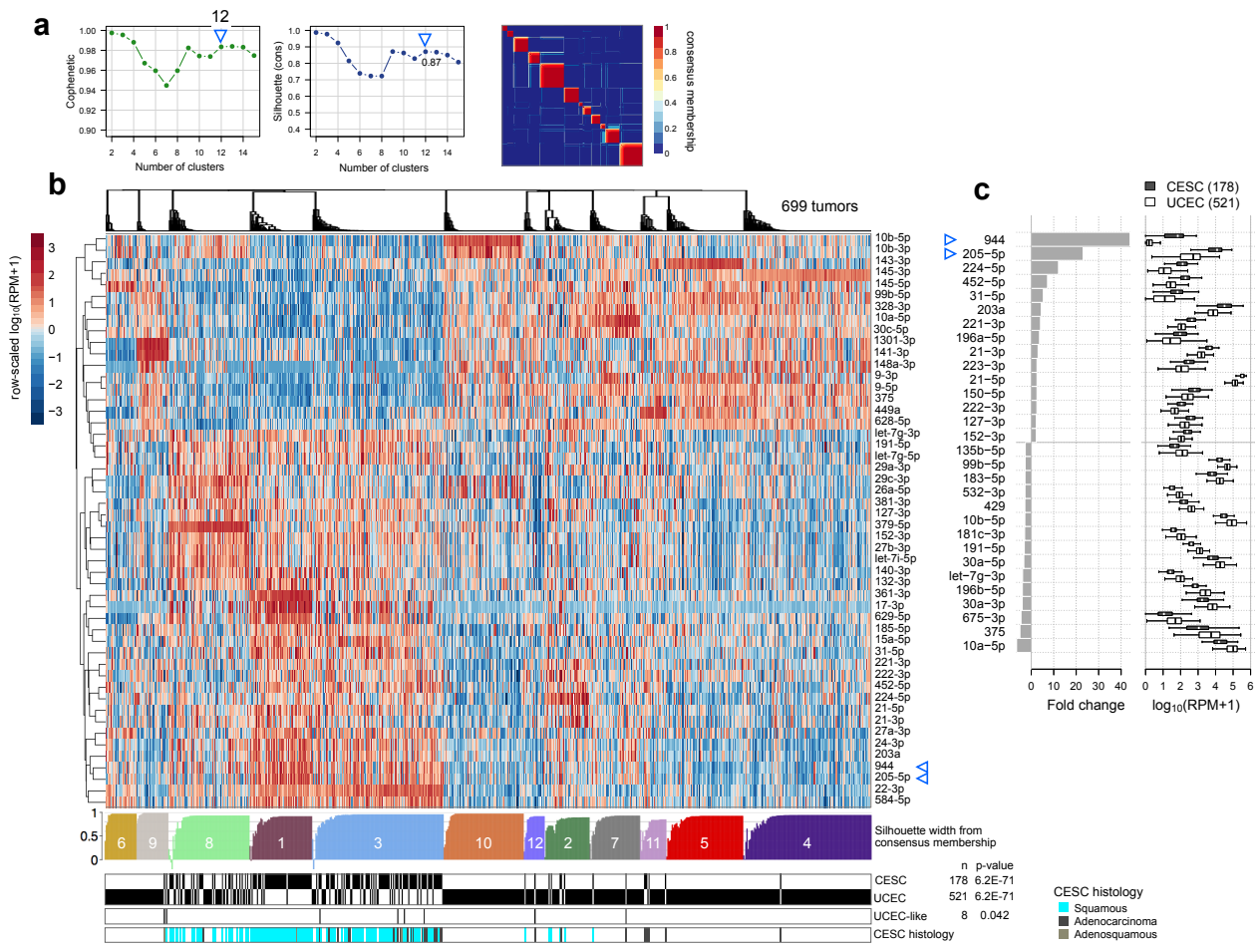
Supplemental Figure S16: miRNAs that may be influenced by DNA methylation. **a,** Covariate tracks showing beta values for DNA methylation and stemloop normalized abundance (RPM). **b,** Distributions across miR-based clusters (all samples) of methylation beta for a correlated DNA methylation probe (above)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

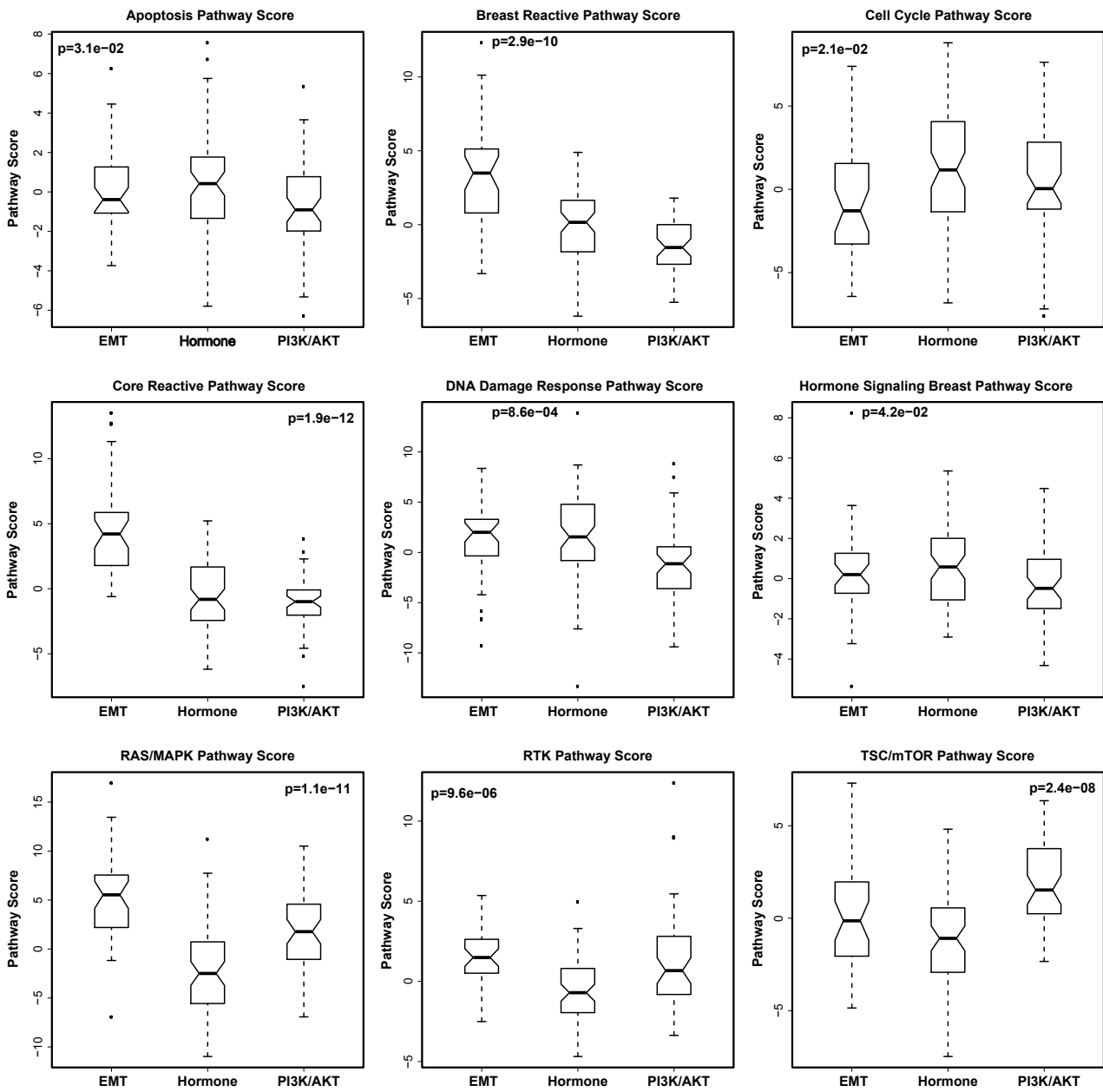


Supplemental Figure S18: Functionally validated potential miR-gene and miR-RPPA targeting for squamous samples. See legend of Supplemental Figure S17.

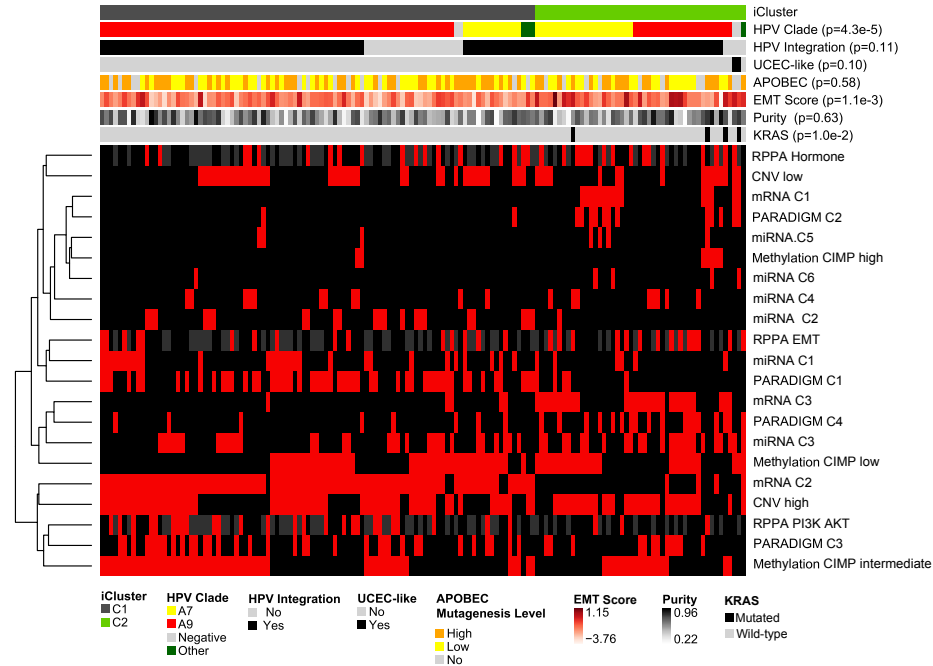


Supplemental Fig. S19

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.

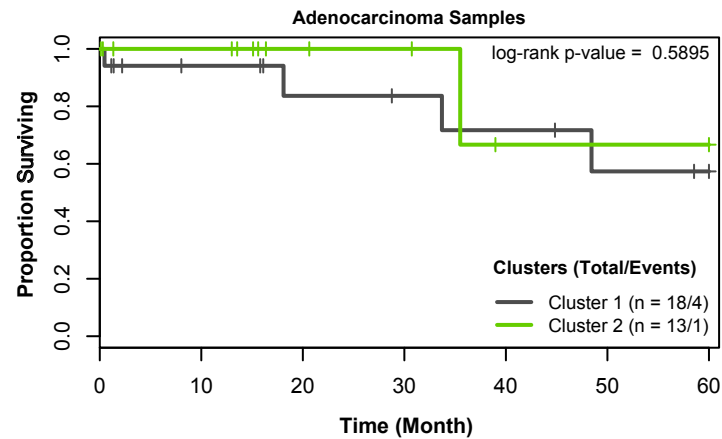
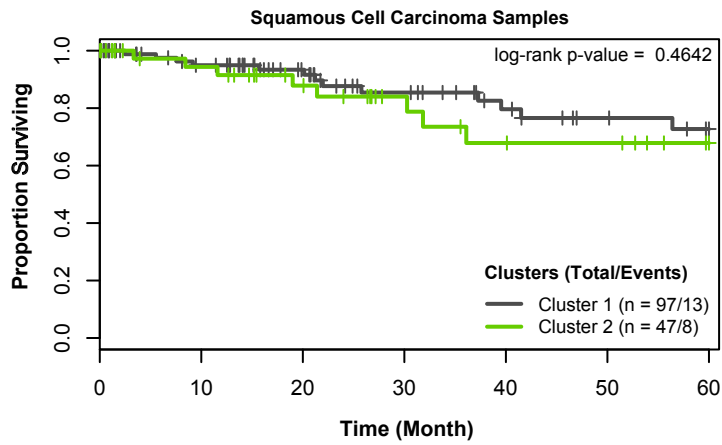
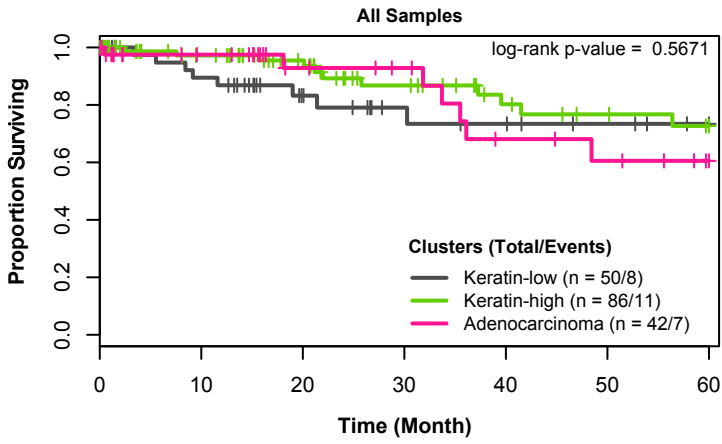


Supplemental Figure S20: Protein signaling pathway scores for all samples. Pathway scores for apoptosis, reactive breast, cell cycle, core reactive, DNA damage response, breast hormone, RAS/MAPK, RTK, and TSC/mTOR signaling pathways are presented with significant pathway score differences between the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

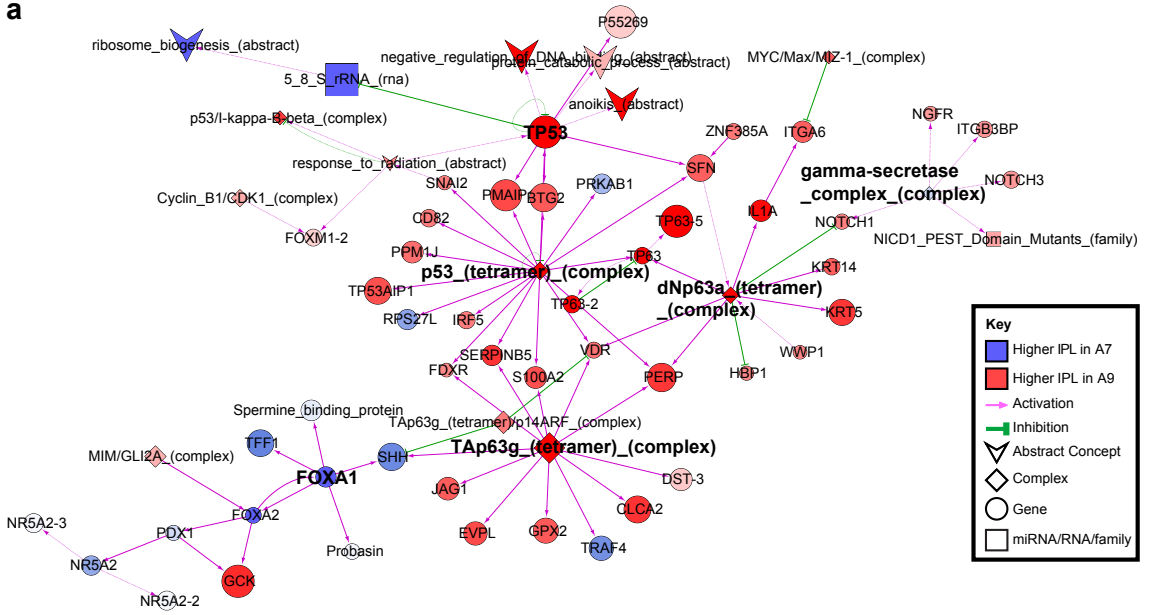
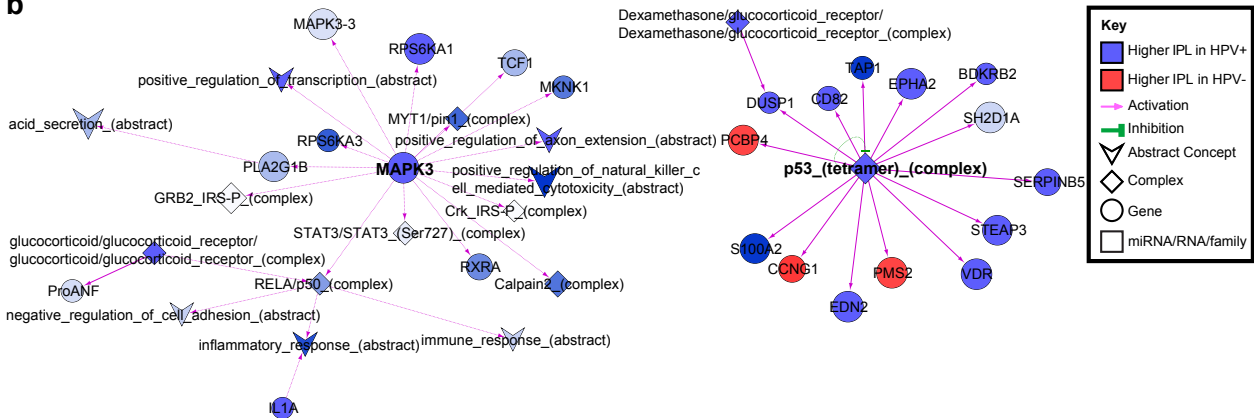
a**Adenocarcinoma****b****Squamous Cell Carcinoma**

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.

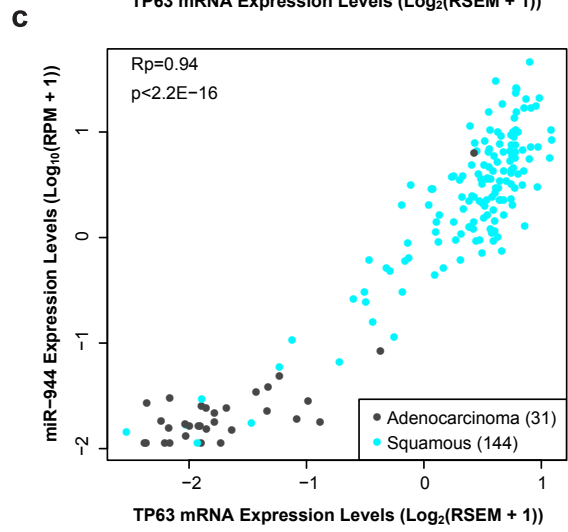
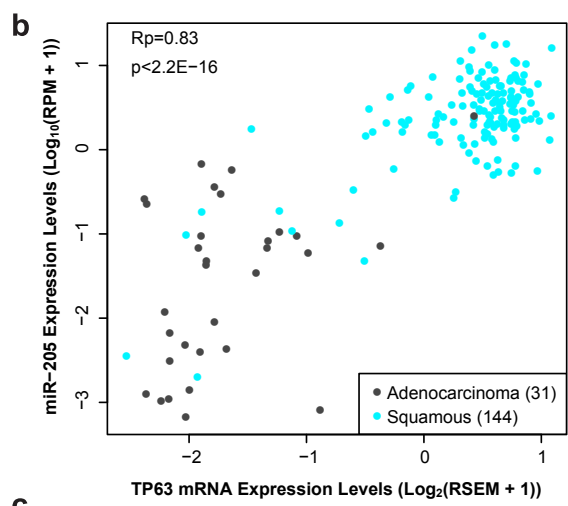
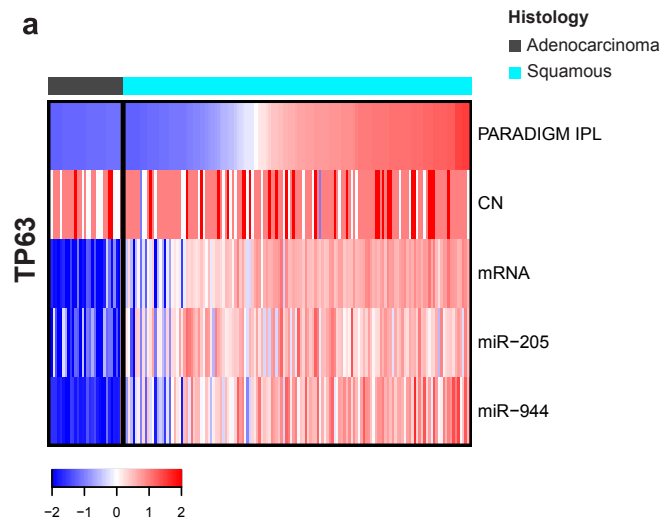
Kaplan Meier Survival Plots



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.

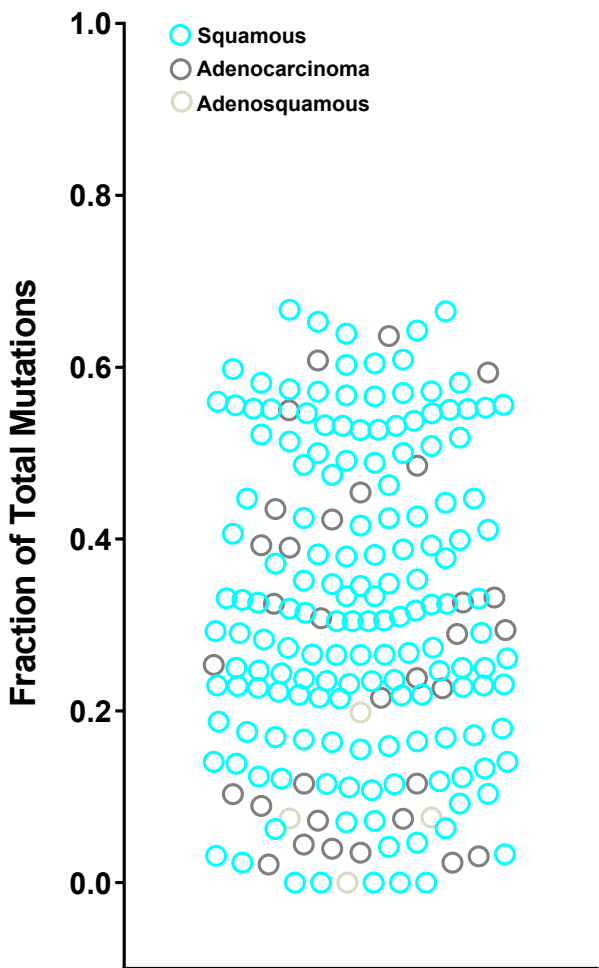
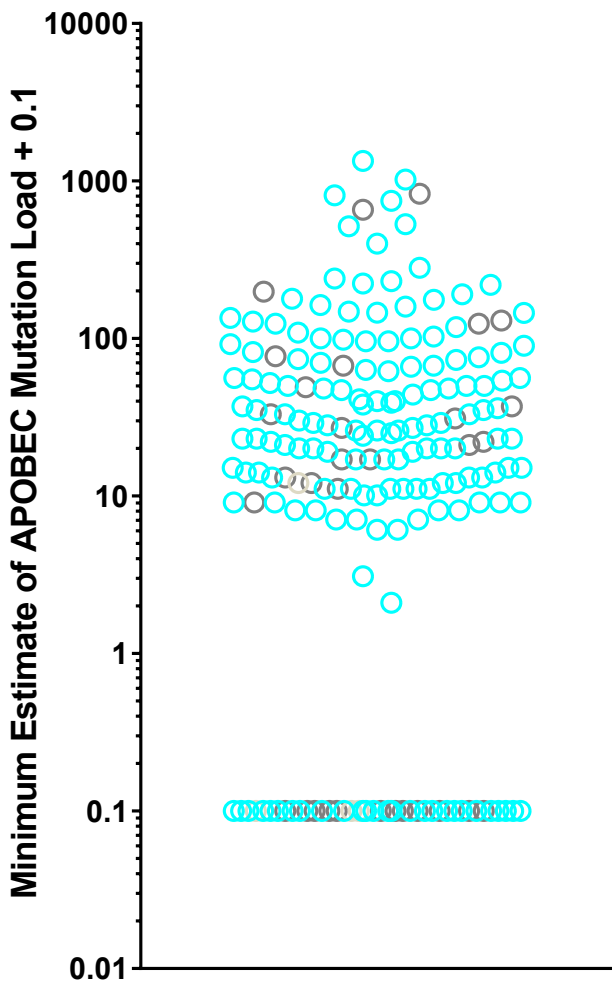
a**b**

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.

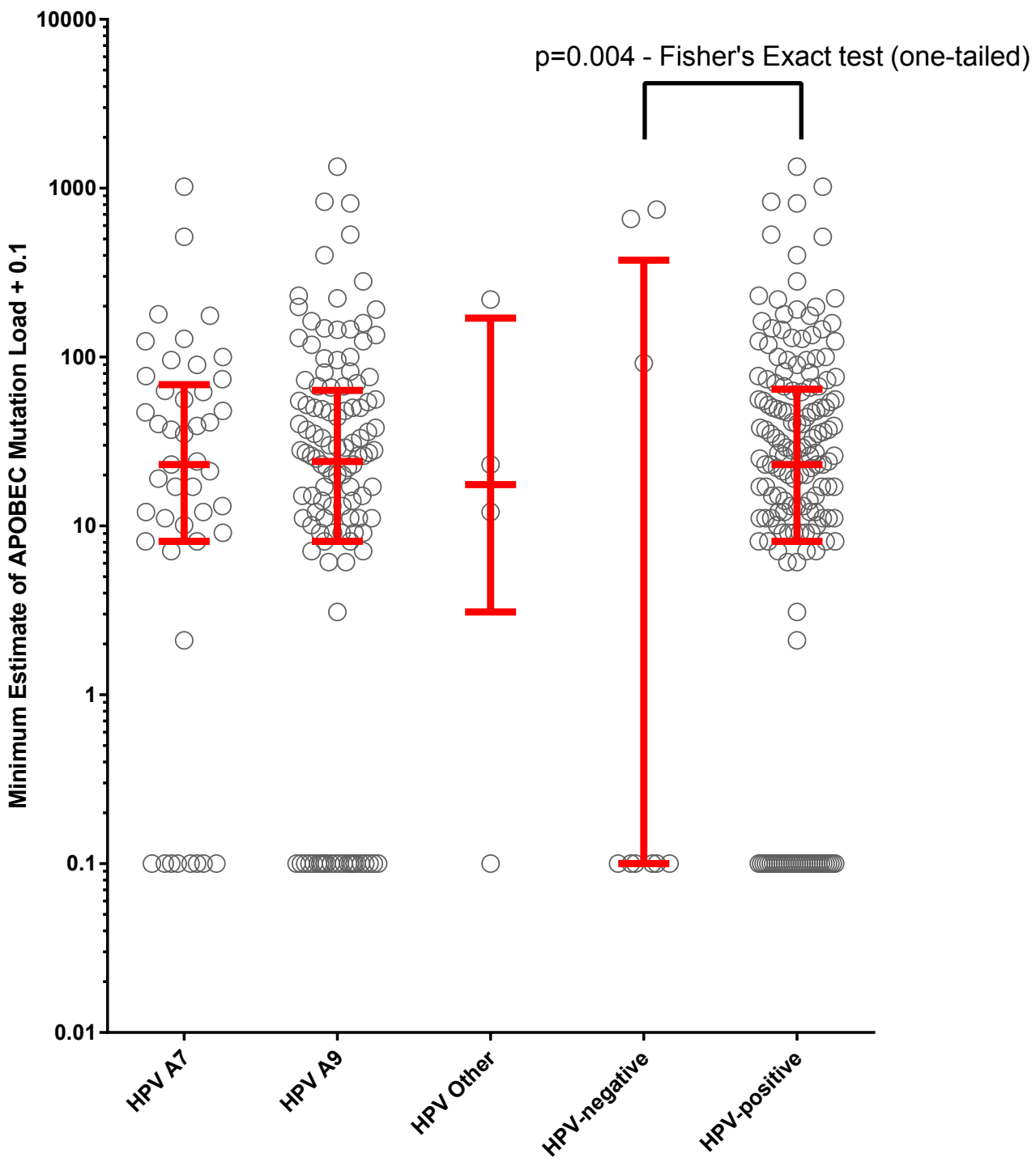


Supplemental Figure S24: p63 as a marker distinguishing squamous carcinomas and adenocarcinomas. **a**, Heatmap showing p63 PARADIGM IPL, copy number (CN), and mRNA expression levels, as well as miR-944 (which is located in the intron of p63) and miR-205 (which has previously been shown to be p63-regulated) 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).

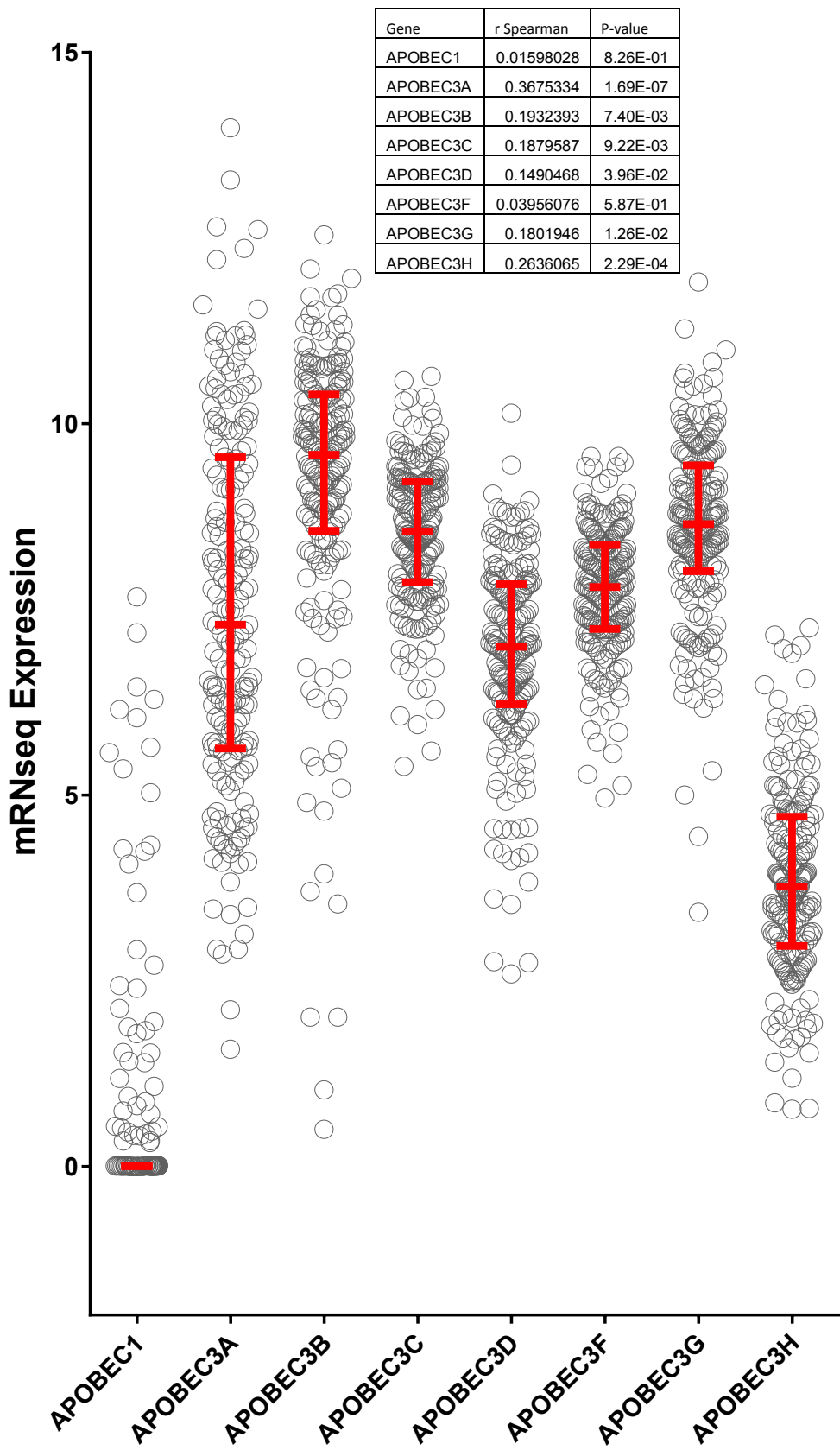
Supplemental Figure S25: FGF signaling in squamous carcinomas relative to adenocarcinomas. **a**, Cytoscape view of interconnected PARADIGM features differentially activated between squamous carcinoma and adenocarcinoma. Node color reflects level of differential activation (red: higher in squamous, blue: 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).

a**b**

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).

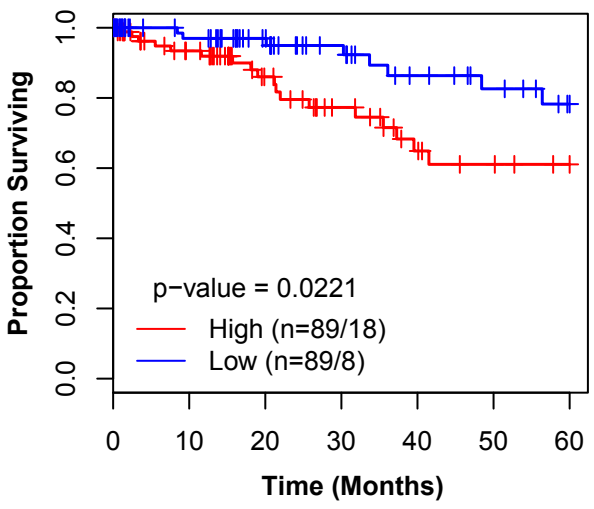


Supplemental Figure S27: APOBEC mutagenesis pattern is reduced in HPV-negative samples compared with HPV-positive samples. APOBEC mutagenesis patterns are presented for samples of different HPV clade (HPV A7, HPV A9, and HPV Other) and status (HPV-negative and HPV-positive).

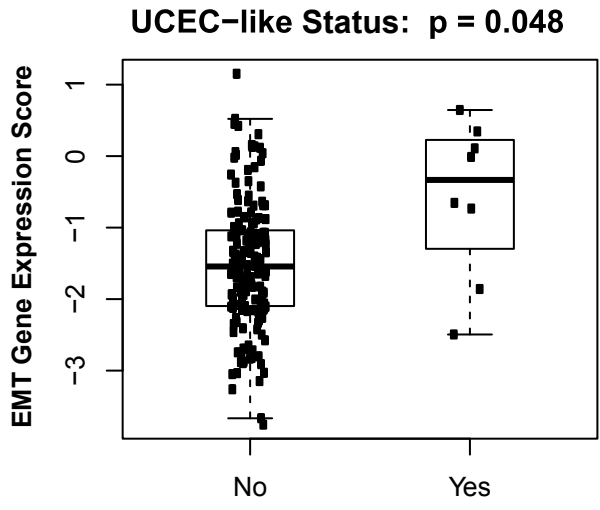


Supplemental Fig. S28: Expression of *APOBEC* genes in TCGA cervical cancer samples correlates with *APOBEC* mutagenesis. mRNA gene expression ($\log_2(\text{RSEM}+1)$) is plotted for *APOBEC1* and *APOBEC3* genes across all 192 Extended Set samples. A small number of zero values for *APOBEC1* expression 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.

a.

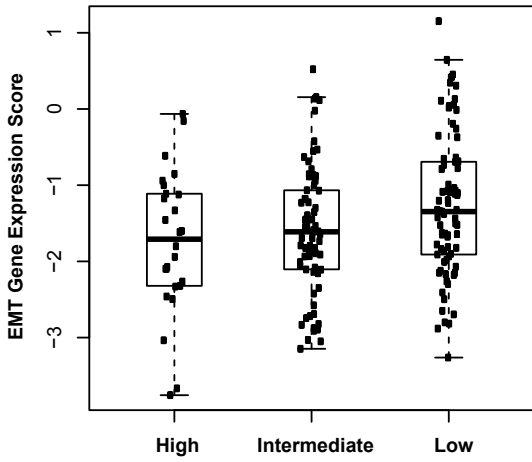


b.

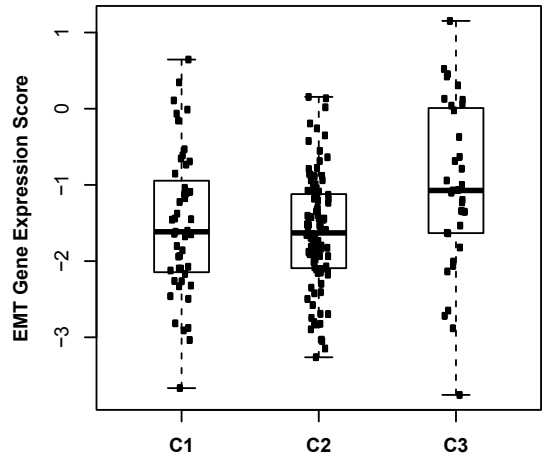


Supplemental Figure S29: EMT mRNA score associations with outcomes and UCEC-like status. **a**, Five year survival analysis comparing the EMT-high vs. EMT-low groups (log-rank $p=0.0221$). **b**, Association of EMT scores with UCEC-like case status (two sample t-test, $p=0.048$).

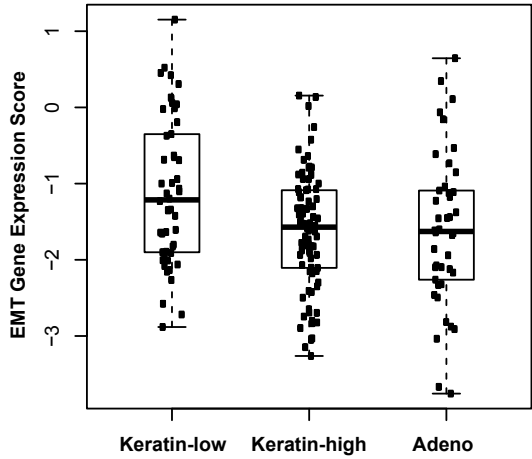
CIMP: $p = 0.024$



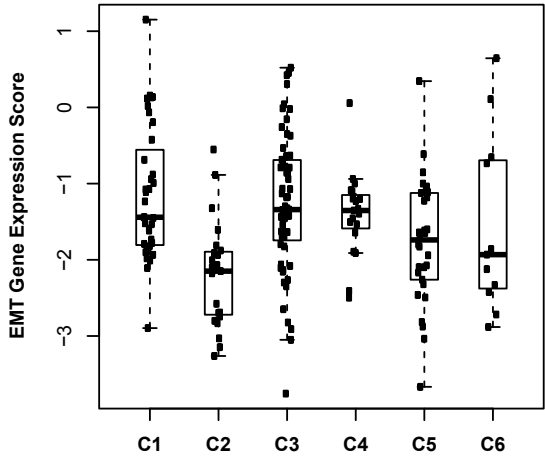
mRNA: $p = 0.003$



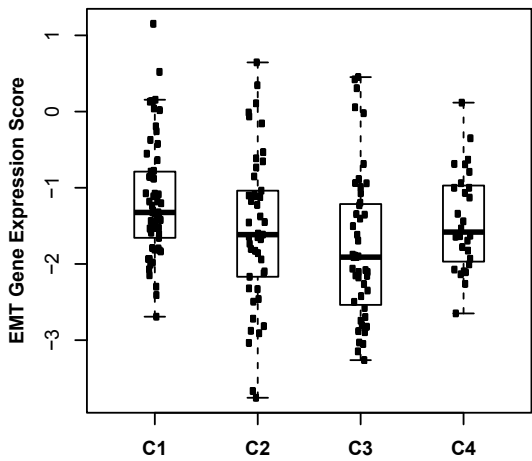
iCluster: $p = 0.003$



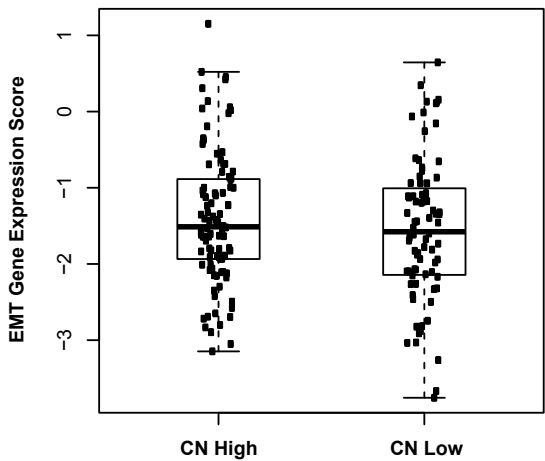
miRNA: $p < 0.001$



PARADIGM: $p = 0.005$



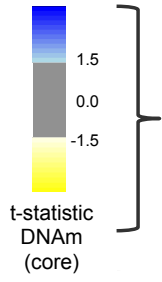
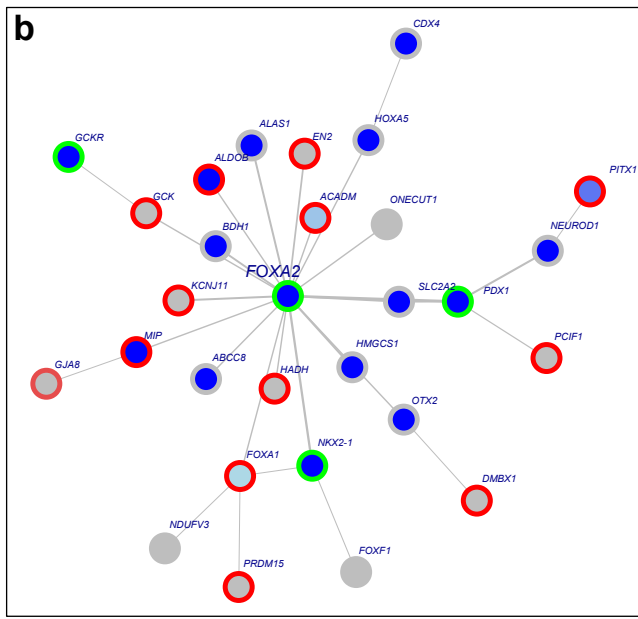
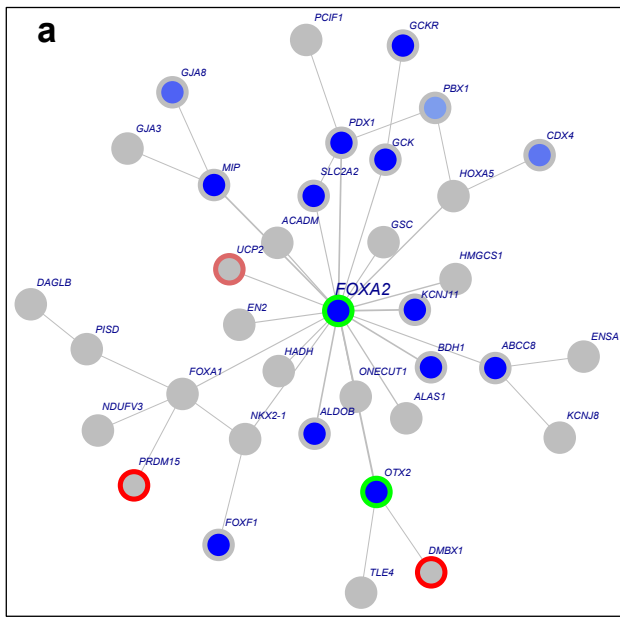
CN: $p = 0.264$



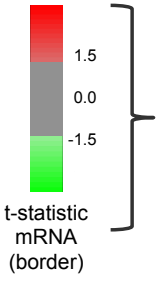
Supplemental Figure S30: Association of EMT mRNA scores with platform clusters.

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


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.



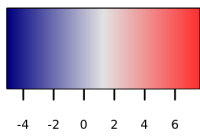
Association between DNAm and HPV status



Association between mRNA and HPV status

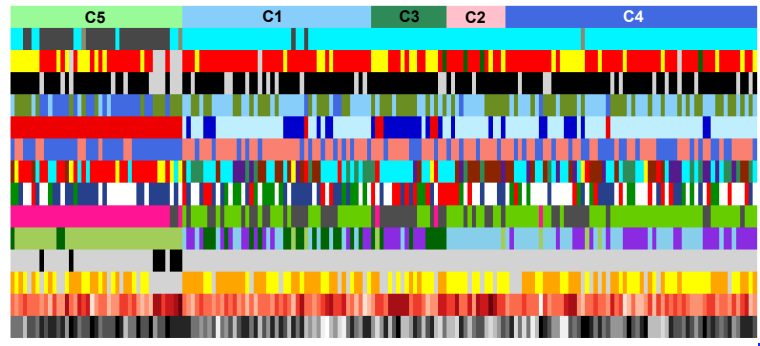
 Increased promoter methylation and decreased expression in HPV positive tumors

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

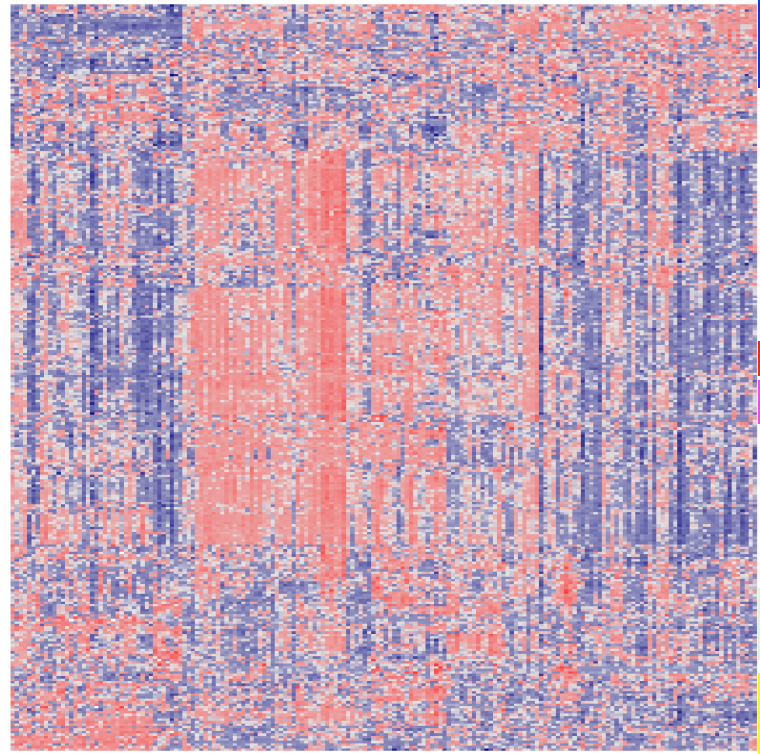
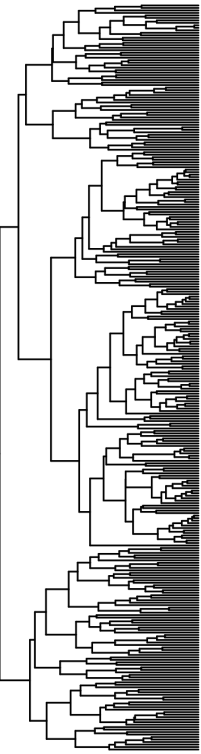


Gene clusters

- Cluster 1 ■
- Cluster 2 ■
- Cluster 3 ■
- Cluster 4 ■
- Cluster 5 ■



- Immune response
- Histology
- HPV clade
- HPV integration
- CIMP call
- mRNAseq
- CN cluster
- miRNA
- RPPA
- iCluster
- PARADIGM
- UCEC-like
- APOBEC mutagenesis level
- EMT score
- Purity



- LTBRAP
- LTBR
- LTBR4H
- GP508
- S105A9
- S105A8
- S105A10
- S105A5
- RTT1
- FOXC1/IG2
- CD27A
- AIM2
- APOL3
- CAC11
- CD313
- CAC19
- CRHR1
- CCR9
- DPS4
- ADORA1
- FOGRT
- NFATC4
- TCF7
- SKAP1
- HEA-2
- SIGIRR
- CHRNA7
- MNK1
- GP84
- HDMC5
- XBP1
- CHST4
- ALOX15
- CCSP2
- ELF3

- 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
 - C2
 - C3
 - C4
 - C5
 - C6

- RPPA cluster
- EMT
 - Hormone
 - PI3K/AKT
 - NA

- iCluster
- Ker-low
 - Ker-high
 - Adeno

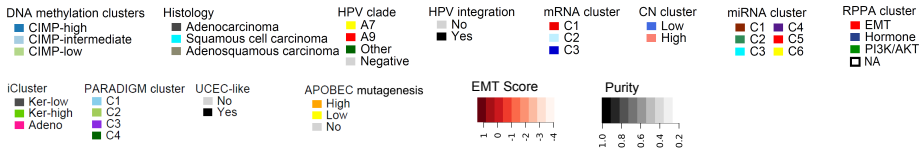
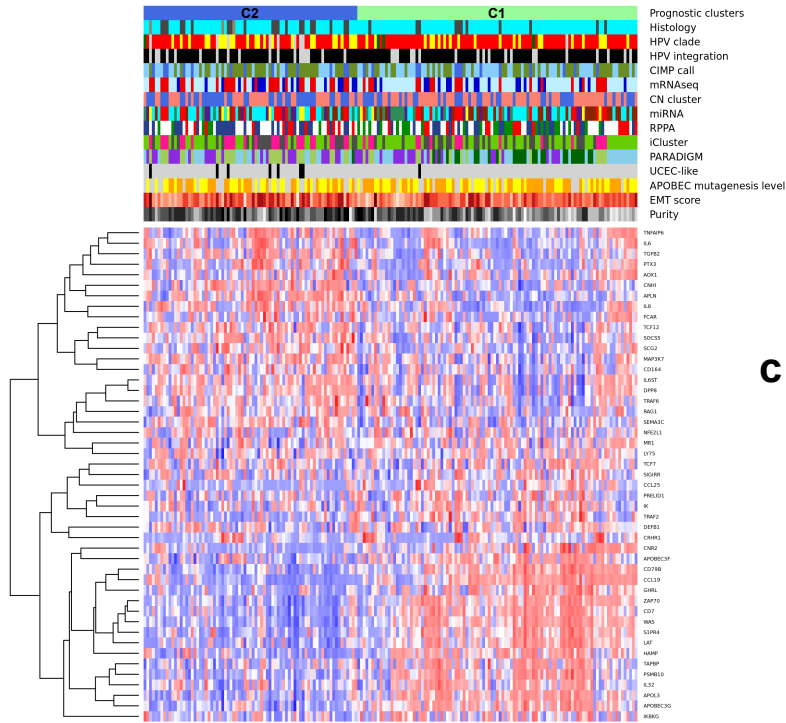
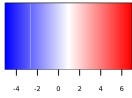
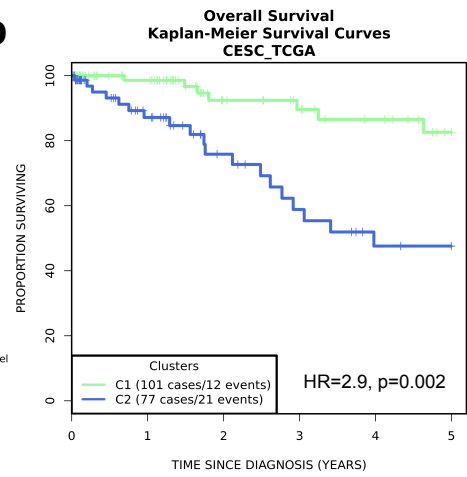
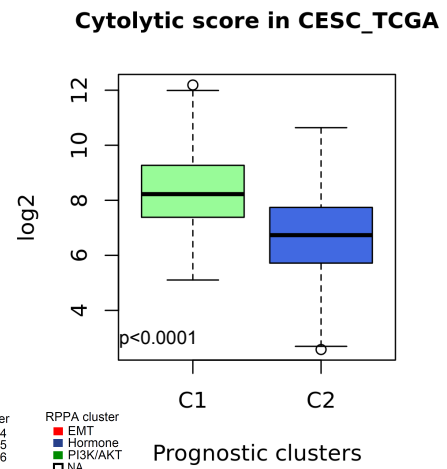
- PARADIGM cluster
- C1
 - C2
 - C3
 - C4

- UCEC-like
- No
 - Yes

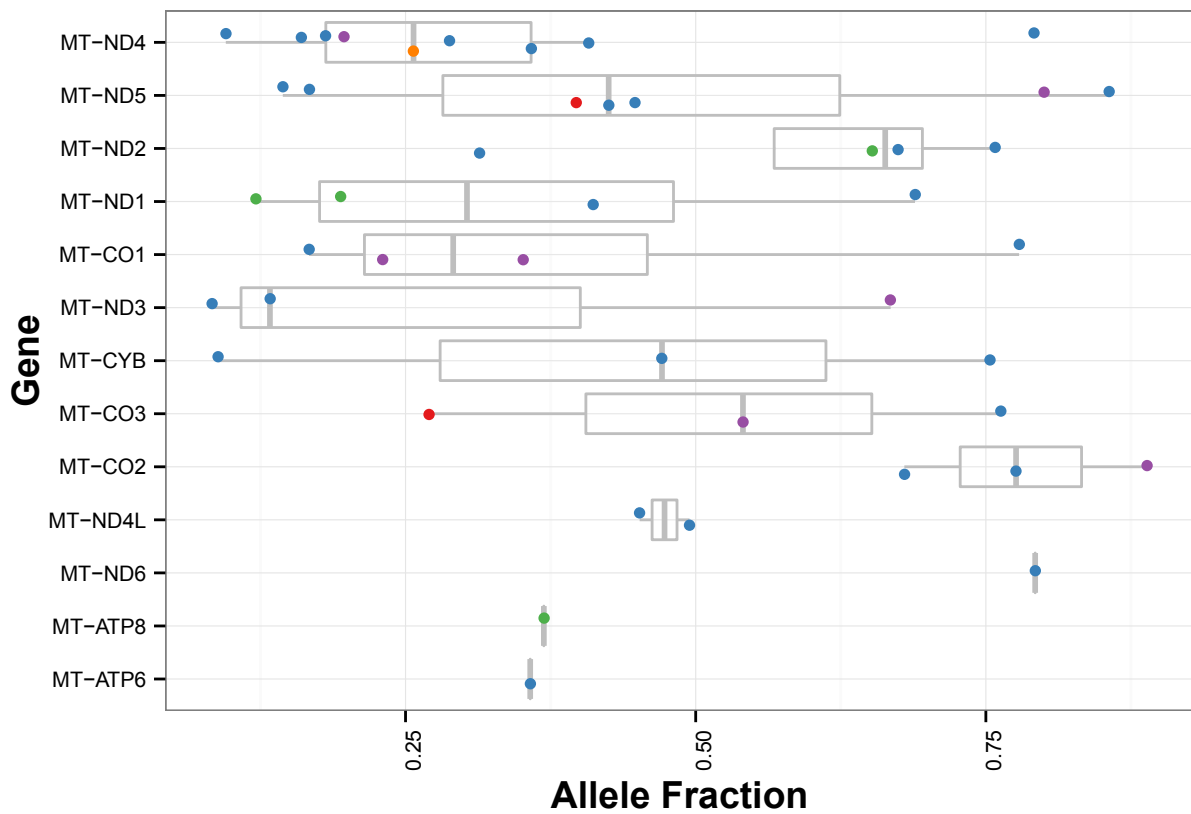
- APOBEC mutagenesis
- High
 - Low
 - No



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.

a**b****c**

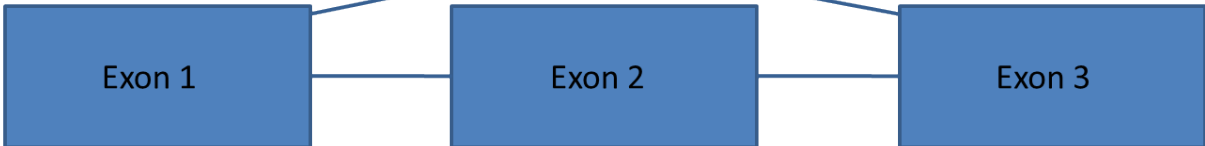
Supplemental Figure S34: Immune response prognostic gene clustering analysis. **a**, Consensus clustering analysis of 178 cervical carcinoma samples based on immune response gene expression selected from ROC analysis. **b**, Kaplan-Meier analysis comparing survival across predicted prognostic clusters.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

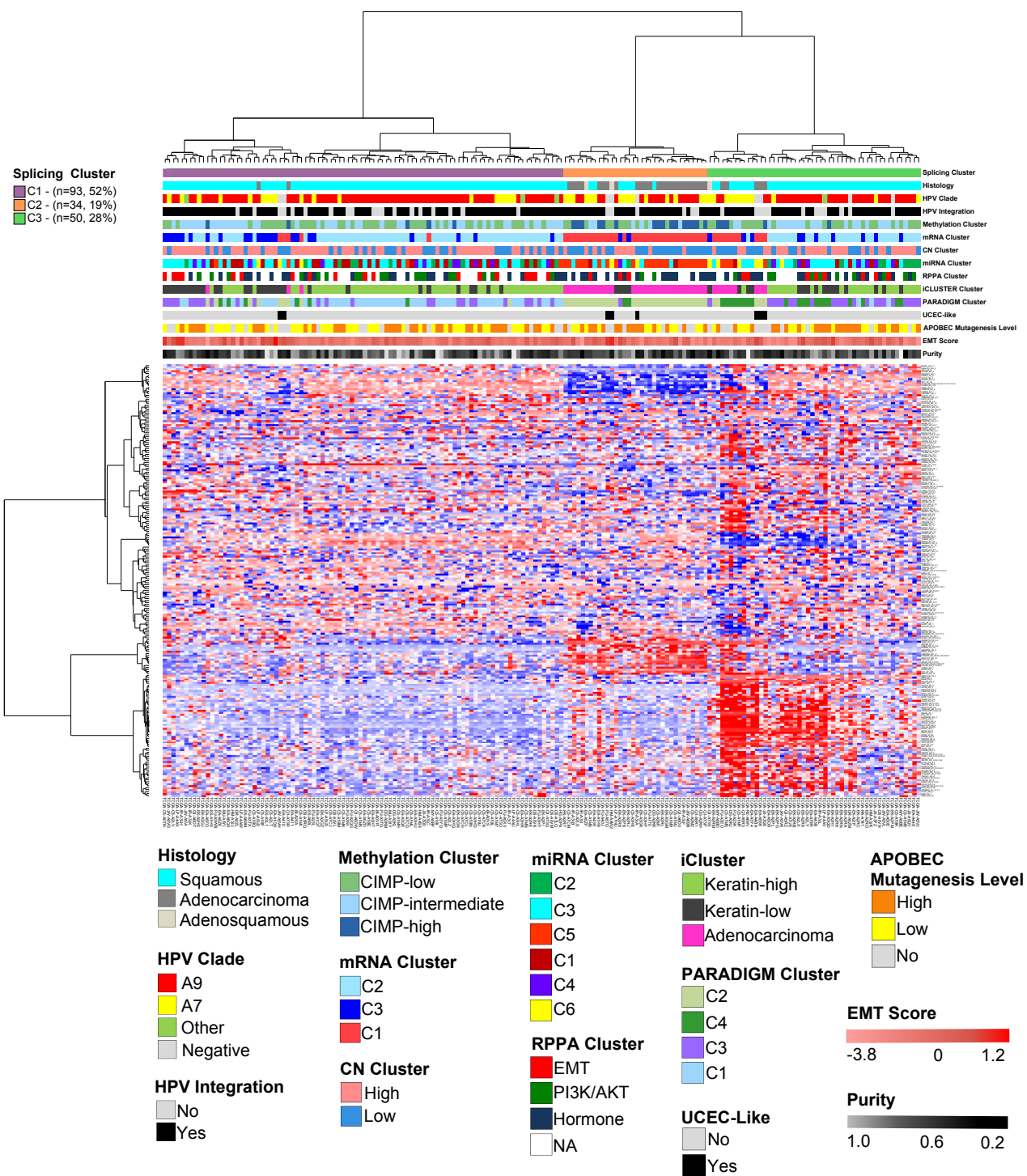
Include Reads



Exclude Reads

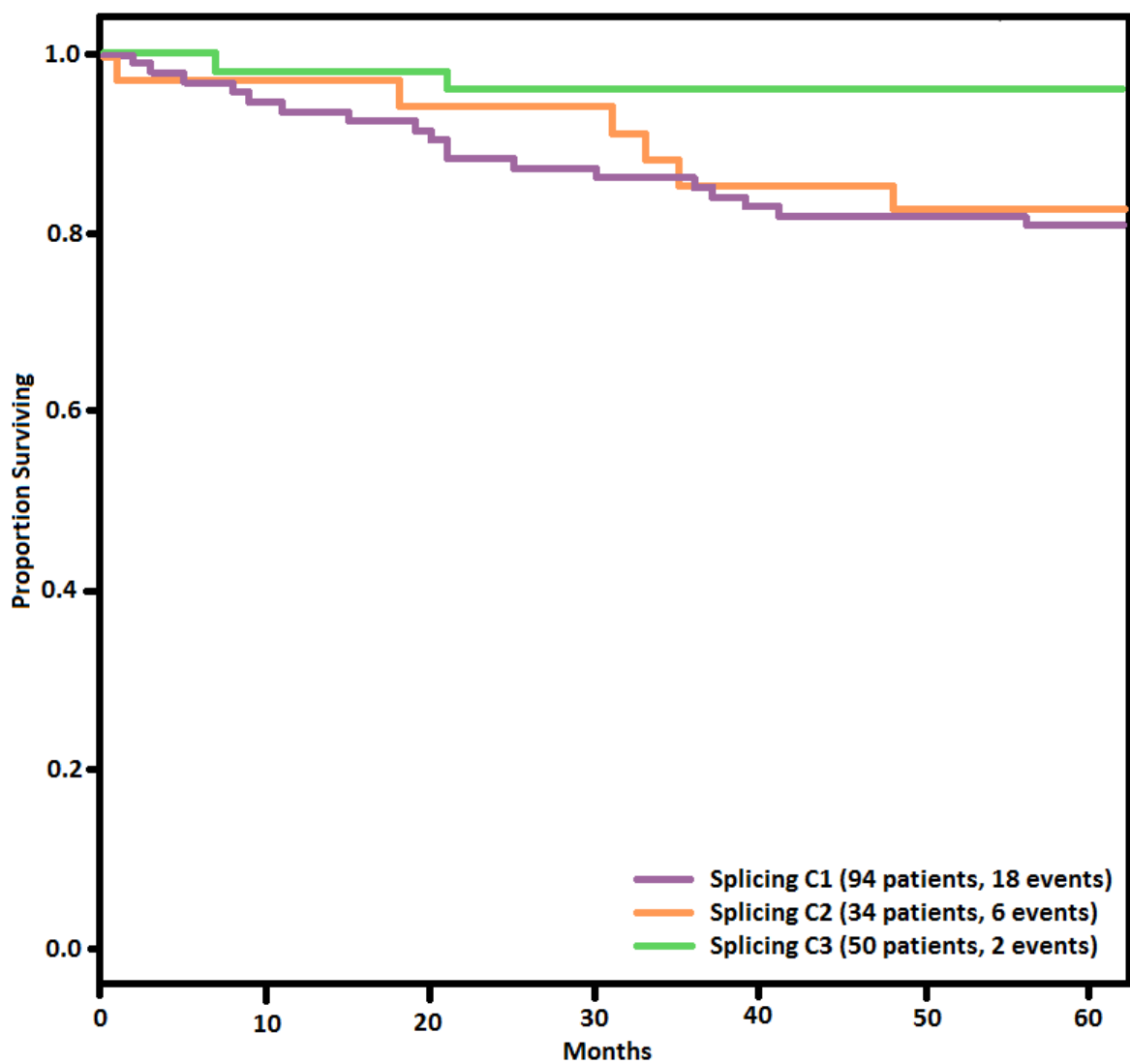


Supplemental Figure S37: PSI as determined from an exon skipping event. The yellow exon and exon junction reads indicate the presence of exon 2 while the red exon 1 – 3 junction reads indicate that exon 2 is spliced out. The PSI is 8 / 10 reads or .8 indicating that ~ 80% of transcripts in the sample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

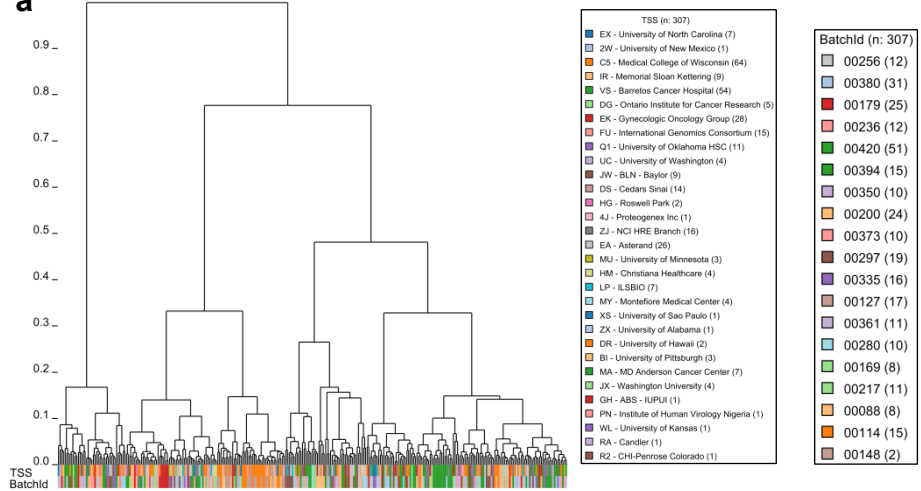
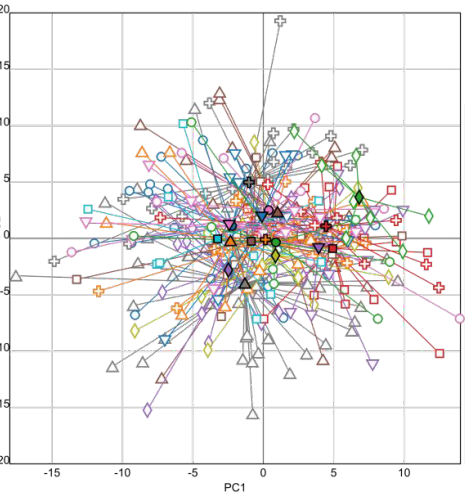
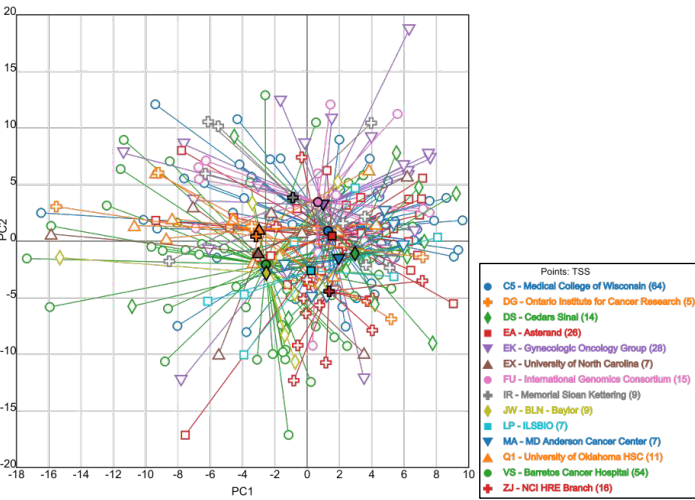


Supplemental Fig. S38

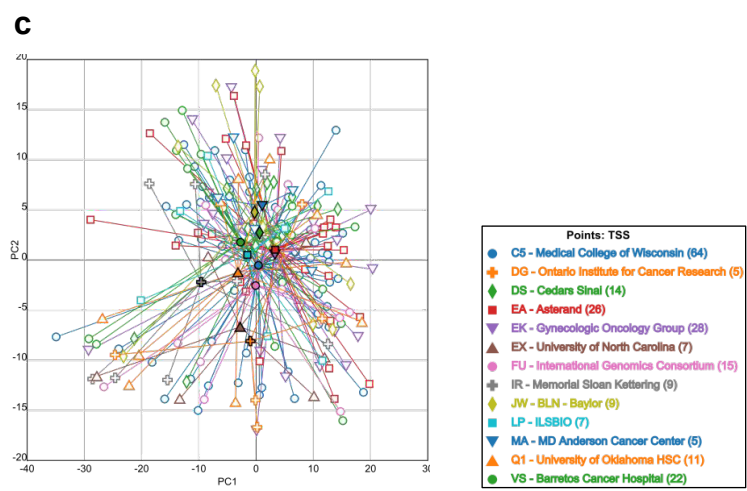
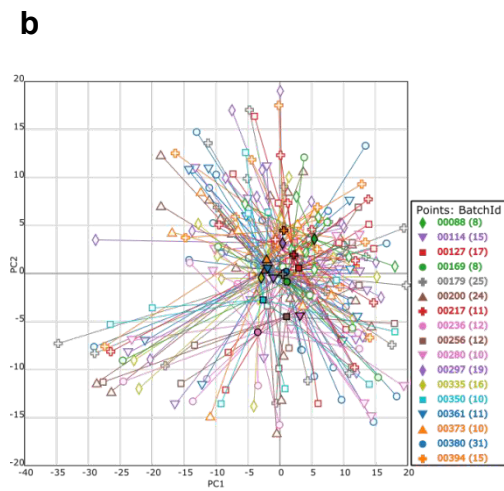
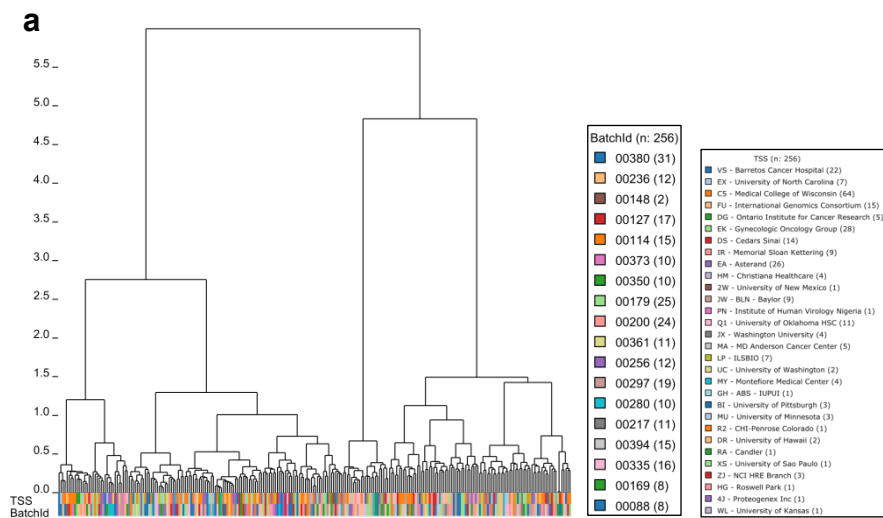
Supplemental Figure S38: Hierarchical clustering heatmap based on splicing events. A hierarchical clustering heatmap is shown of mean-centered PSI values for high variation splice events in samples.



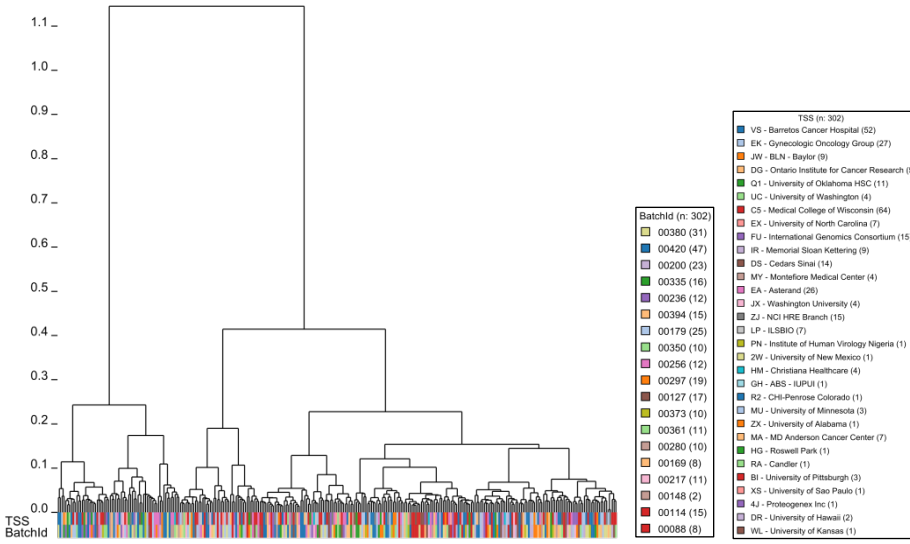
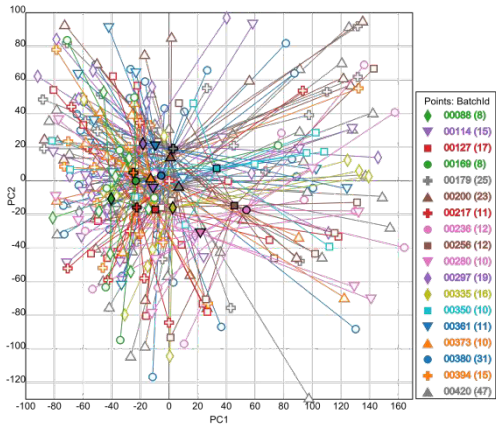
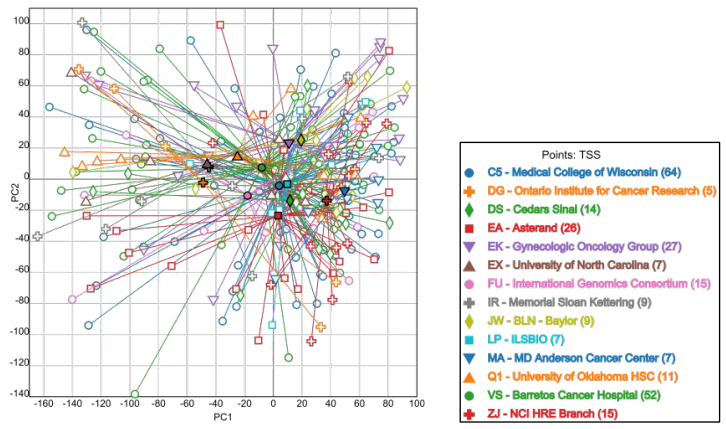
Supplemental Figure S39: Survival analysis for the three splicing clusters. Kaplan-Meier analysis was performed to assess survival differences across all three splicing clusters.

a**b****c**

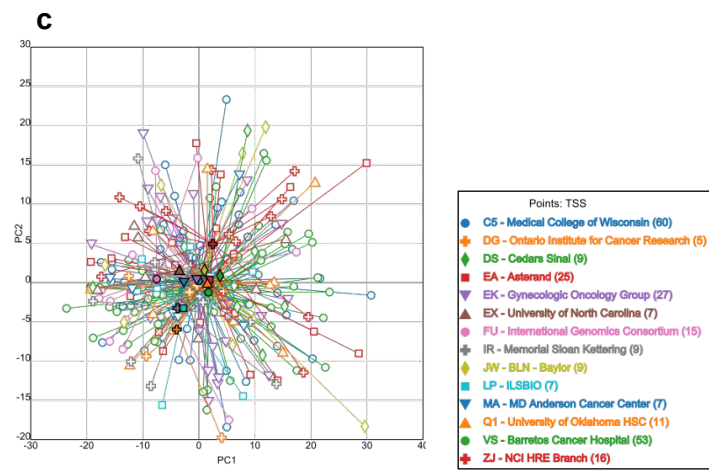
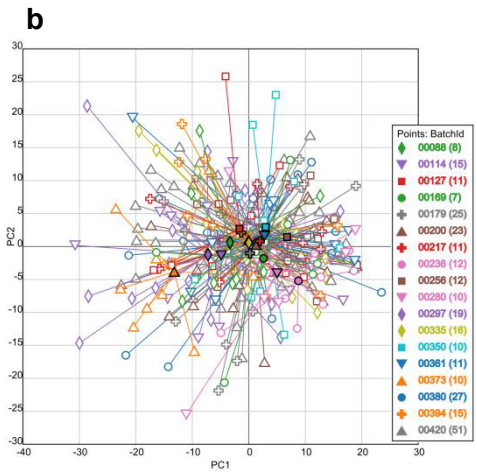
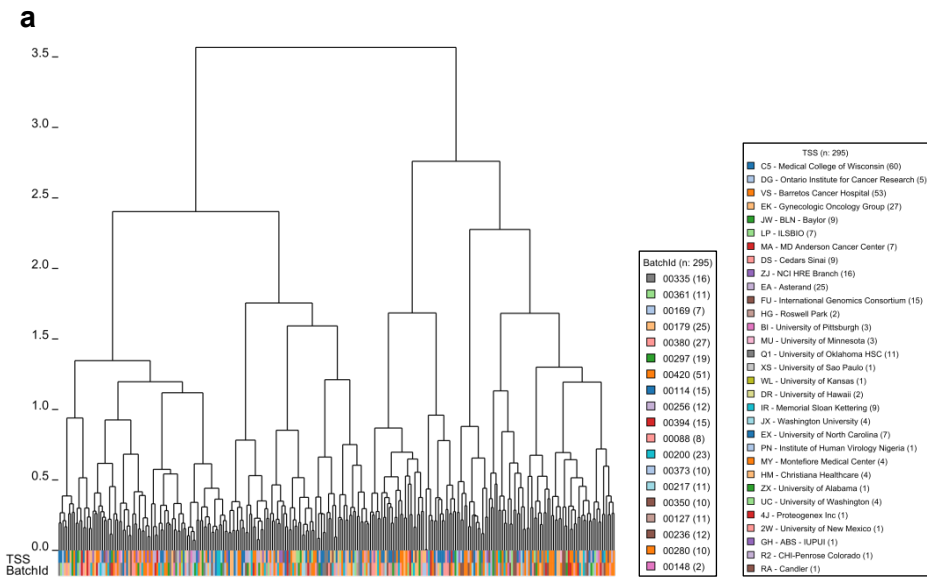
Supplemental Figure S40: miRNA batch effects. **a**, Hierarchical clustering of samples based on miRNA sequencing data. **b**, PCA for miRNA expression from miRNAseq data, with samples connected by centroids according to batch ID. **c**, PCA for miRNA expression from miRNAseq data, with samples connected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



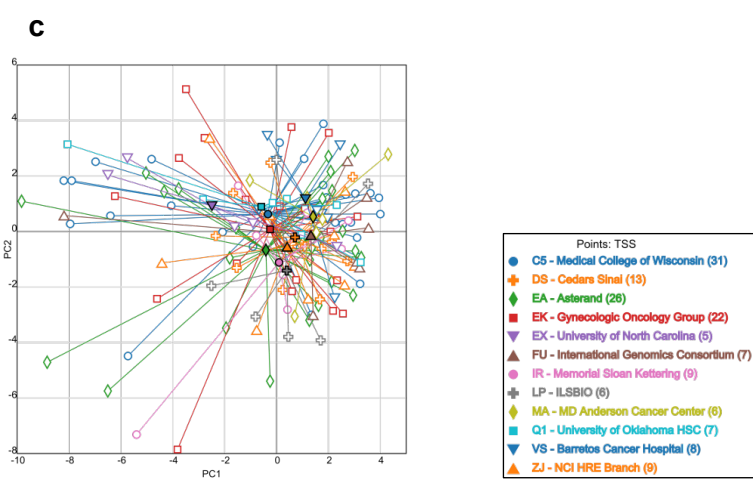
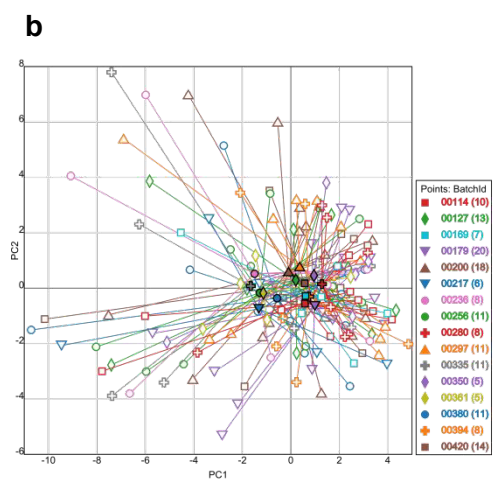
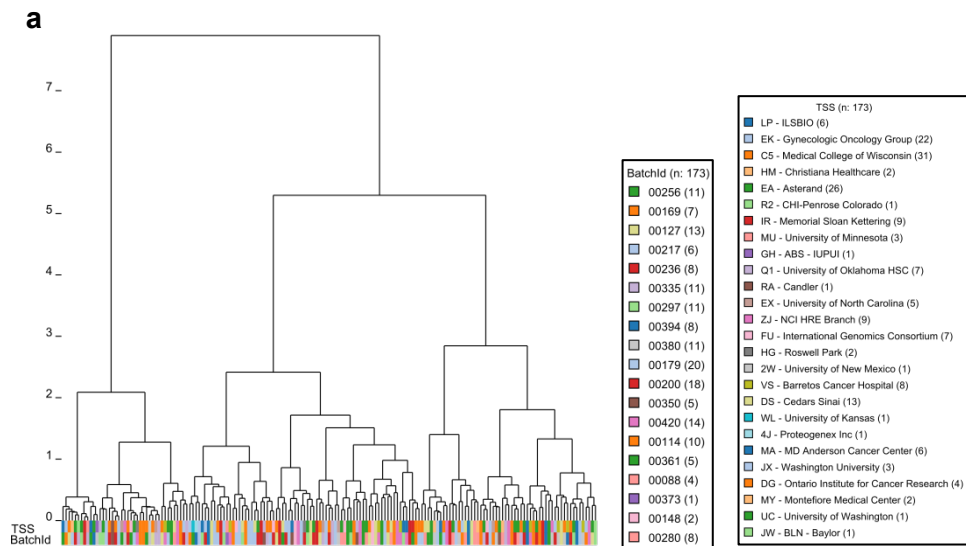
Supplemental Figure S41: DNA methylation batch effects. **a**, Hierarchical clustering of samples based on DNA methylation data. **b**, PCA for DNA methylation, with samples connected by centroids according to batch ID. **c**, PCA for DNA methylation, with samples connected by centroids according to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

a**b****c**

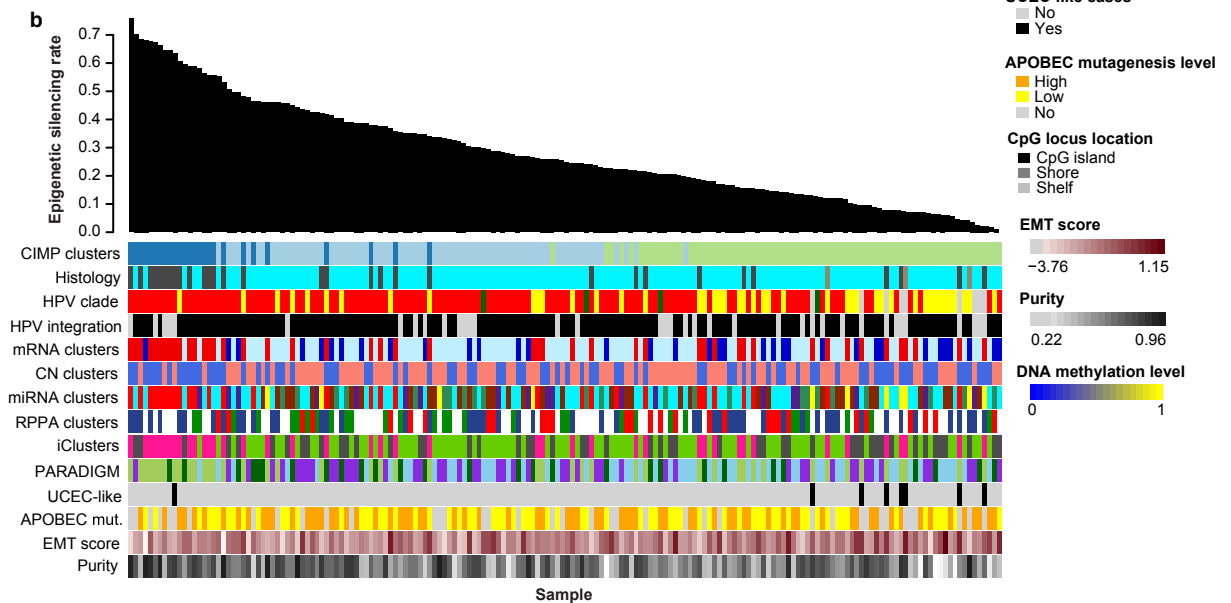
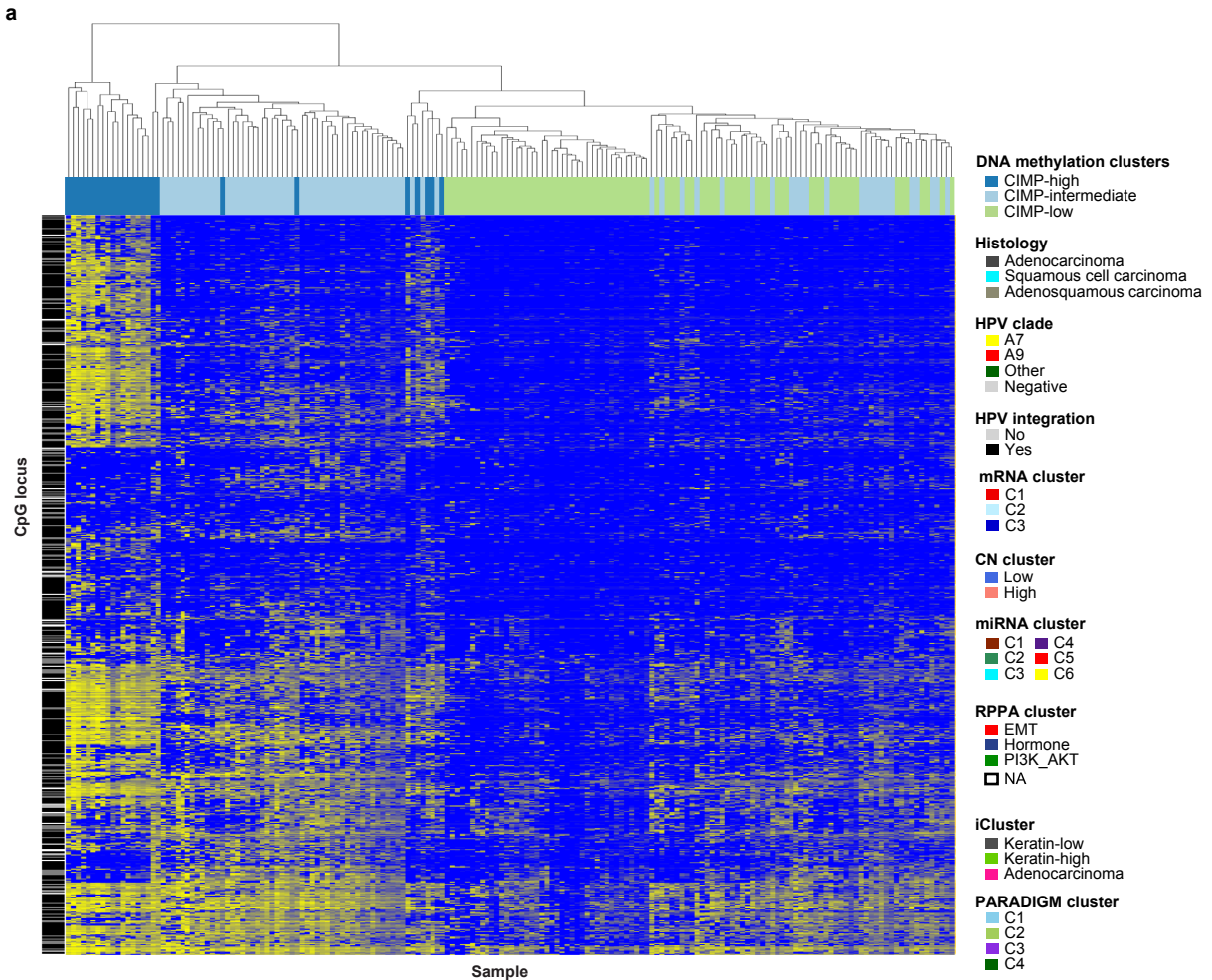
Supplemental Figure S42: RNAseq batch effects. **a**, Hierarchical clustering of samples based on RNAseq expression data. **b**, PCA for RNAseq, with samples connected by centroids according to batch ID. **c**, PCA for RNAseq, with samples connected by centroids according to TSS.



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.



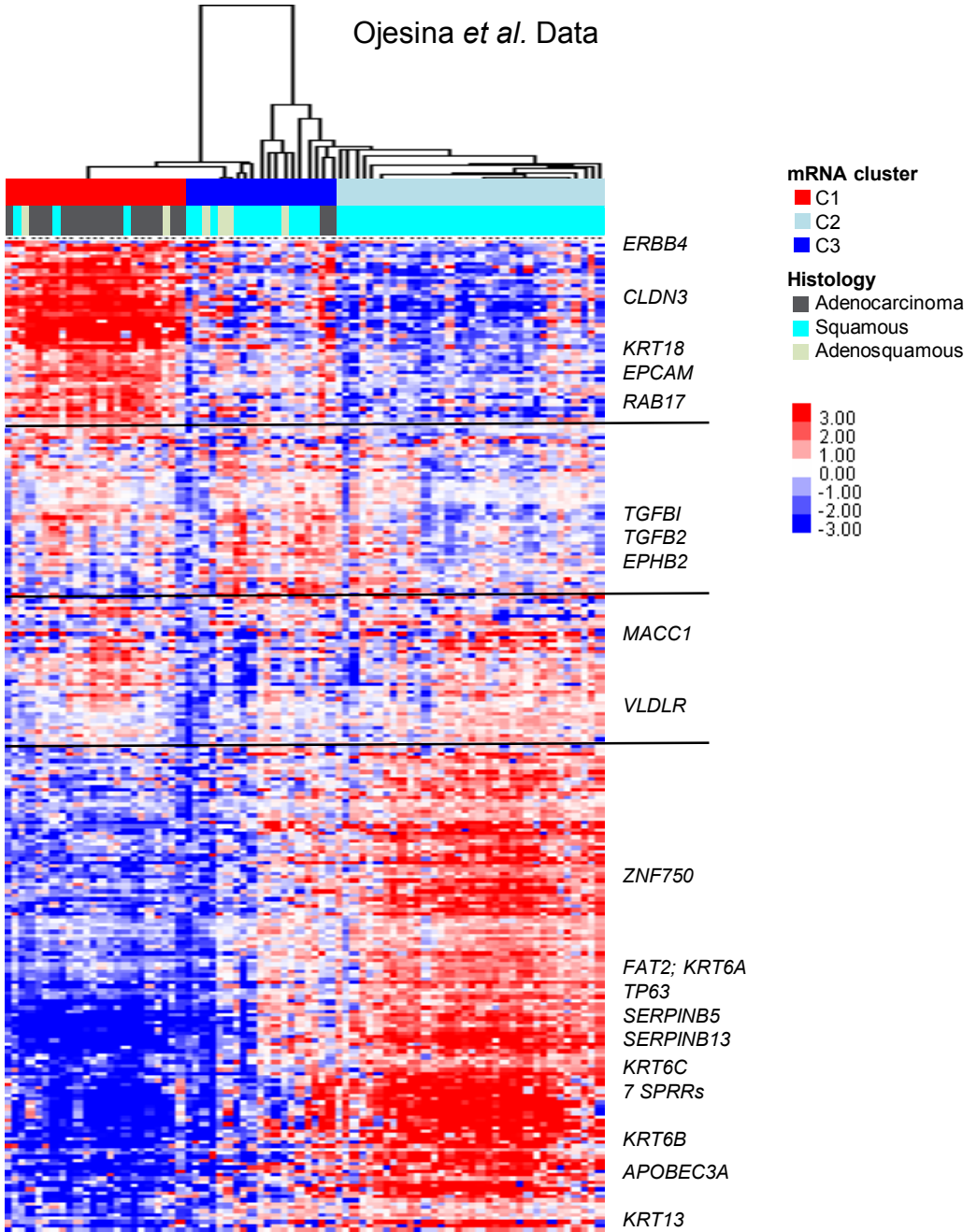
Supplemental Figure S44: RPPA batch effects. **a**, Hierarchical clustering of samples based on RPPA data. **b**, PCA for RPPA data, with samples connected by centroids according to batch ID. **c**, PCA for RPPA data, with samples connected by centroids according to TSS.



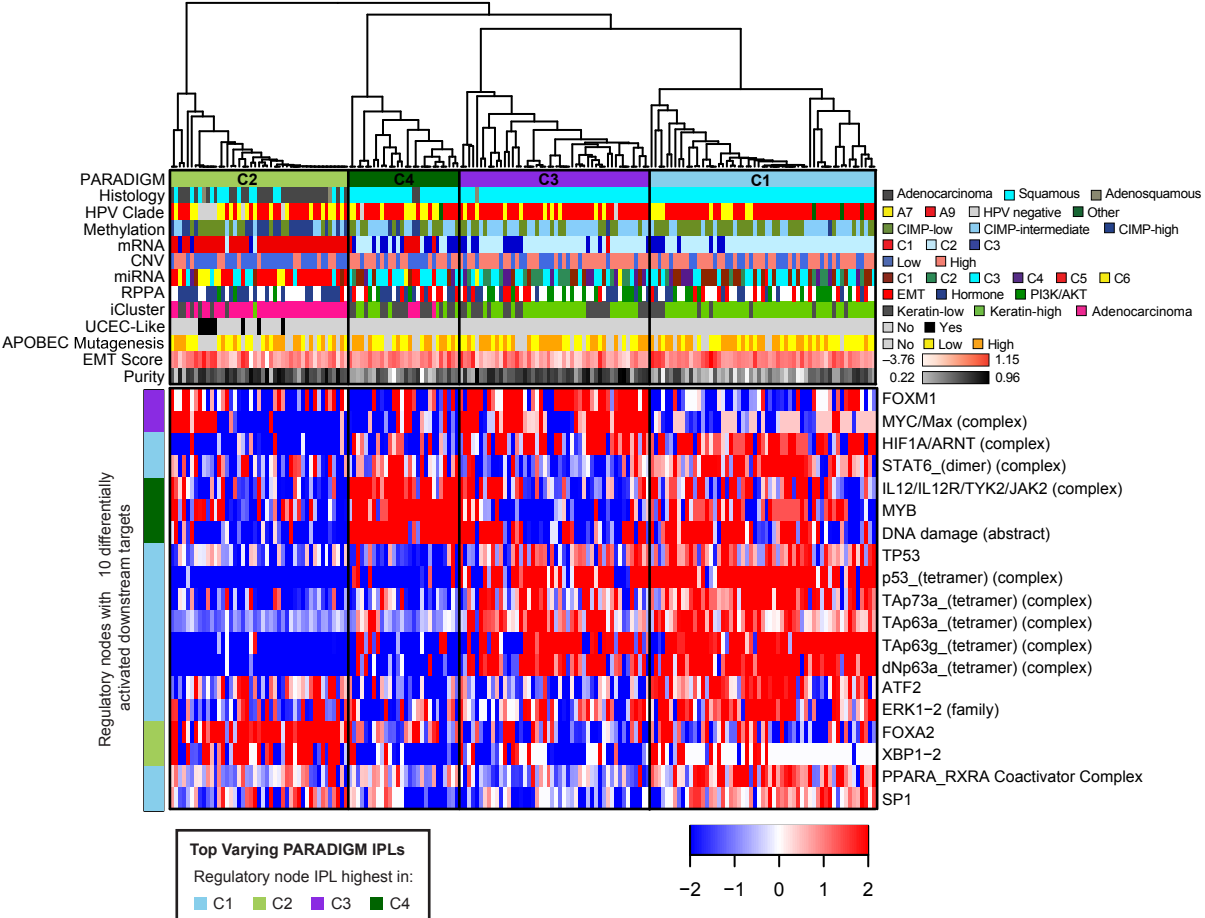
Supplemental Figure S45: Epigenetically silenced gene frequencies. **a**, Heatmap showing unsupervised hierarchical clustering of beta values of 178 samples and epigenetically silenced probes from Supplemental Table 12. Samples are presented in columns and CpG loci are presented in rows. An annotation 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.

Ojesina *et al.* Data



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.