# Landscape of Genomic Alterations in Cervical Carcinomas

***Ojesina AI, Lichtenstein L, et al.***

## SUPPLEMENTARY INFORMATION

# SUPPLEMENTARY NOTES

## *SN1. Sample Collection*

Tumors were prospectively collected serially over a period of 10 years in Norway and 3 months in Mexico. Based on the inclusion criteria set for the study, we only analyzed high quality DNA extracted from fresh frozen tissues with high tumor content, obtained from patients consented for genomic studies.

### 1A. Norwegian Samples

Samples in the Norwegian cohort (N = 103 tumor/normal pairs) were collected in a population based setting from consented patients treated at the Department of Obstetrics and Gynecology, Section of Gynecological Cancer, Haukeland University Hospital, Bergen, Norway, from May 2001 to May 2011. The study site is a referral hospital for all patients with cervical cancer from Hordaland representing approximately 10% of the Norwegian population. Hordaland has a similar incidence rate and prognosis for cervical carcinoma as the total Norwegian population[1]. Patient characteristics are listed in Supplementary Table 2. Surgically resected tumors or biopsies were freshly frozen in nitrogen and stored at minus 80°C. In order to prevent contamination between specimens, the microtome was subjected to brush wipe down followed by alcohol disinfection between specimens. The specimen was evaluated for tumor purity by cutting one section for frozen section before restoring the tissue. If the specimen qualified for inclusion, the tissue was further processed, first for DNA extraction and subsequently RNA extraction if sufficient tissue was available.

Genomic DNA and RNA were extracted from tumors found by frozen section investigations to have > 50% malignant epithelial component (median 80%). In line with these characteristics, the sample collection investigated by deep sequencing was enriched for stage IB surgically resectable, grade III tumors, and more often in need of adjuvant therapy compared to the not sequenced FIGO stage I tumors. Otherwise, there were no other significant differences between the patient cohort and other patients treated in the same period, including disease specific survival. Blood samples were used for extraction of normal DNA. These investigations were approved by the Norwegian Social Science Data Services (15501) and the local Institutional Review Board (REKIII nr. 052.01). Note that 3 tumor/normal pairs failed quality control and were censored, hence we report data on only 100 pairs.

### 1B. Mexican Samples

All Mexican samples (N=16 tumor/normal pairs) were collected from Central Mexico (Instituto Mexicano del Seguro Social project R-2012-785-016). Following local legislation, an additional IRB approval by the Comision Federal para la Proteccion contra Riesgos Sanitarios (COFEPRIS) was obtained. Cervical cancer specimens were paired with peripheral white blood cells. Tissue samples were snap-frozen in liquid nitrogen until DNA extraction, and blood was stored at 4 C for 24 h before processing. Afterwards, pathological review on representative sections of the tumor samples was performed by two independent pathologists. A third pathologist was consulted when consensus for a particular sample was needed. A representative fragment was cut from each tissue using a new surgical blade (DLP Surgical Blades, Dentilab, Mexico). Care was taken to avoid contamination among specimens by processing each sample separately with new and disposable materials. Note that 1 tumor/normal pair failed quality control and were censored, hence we report data on only 15 pairs

## *SN2. Nucleic Acid Extraction*

### 2A. DNA and RNA Extraction from Norwegian Samples

DNA from primary cervical carcinoma lesions was extracted from freshly frozen biopsies. DNA was isolated by digestion over night at 65°C in lysis buffer containing proteinase K, followed by a standard ethanol precipitation. DNA from blood was extracted using a standard Qiagen DNA extraction kit according to manufacturer's protocol. RNA was extracted using the Qiagen RNeasy Mini Kit according to manufacturer's protocol.

### 2B. DNA Extraction from Mexican Samples

DNA was extracted using commercial kits (QIAamp DNA kit, Qiagen), following the manufacturer´s instructions. Only tissues having more than 40% tumoral cellularity were used for downstream analyses. No RNA was extracted for the Mexican samples.

## *SN3. HPV Typing*

### 3A. DNA-based HPV typing

Our primary methods for determining HPV status were two HPV DNA-based PCR assays:

1. A multiplex flourescent-PCR kit that targets the E6 and E7 regions of 13 high-risk HPV (Genomed f-HPV; **http://www.f-hpv.com/index.html**) using manufacturer's instructions. This method is comparable to the Digene HC test[2] and has the additional advantage of being able to identify specific HPV types.
2. An HPV PCR-MassArray method using real-time competitive polymerase chain reaction and matrix-assisted laser desorption/ionization-time of flight mass spectroscopy with separation of products on a matrix-loaded silicon chip array[3,4]. Multiplex PCR amplification of the E6 region of 16 discrete high-risk HPV types (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73 and 90), 2 low-risk HPV types (HPV 6 and 11) and human GAPDH control was run to saturation followed by shrimp alkaline phosphatase quenching. Amplification reactions included a competitor oligo identical to each natural amplicon except for a single nucleotide difference. Probes that identify unique sequences in the oncogenic E6 region of each type were used in multiplex single base extension reactions extending at the single base difference between wild-type and competitor HPV so that each HPV type and its competitor were distinguished by mass when analyzed on the MALDI-TOF mass spectrometer.

As depicted in Supplementary Table 14 and Supplementary Fig.19, 109 (96%) of 113 tumors were found to be positive by the f-HPV test, while 102 (91%) of 112 tumors were found to be HPV-positive by the PCR-MassArray method. Of the 111 tumors with data available from both tests, there was agreement in 101 (91%) tumors for HPV positivity.

Due to the lack of perfect agreement between the 2 tests (and the low frequency of tumors called as HPV-negative by both methods), we have not performed any comparative analysis on HPV-positive versus HPV-negative tumors.

## 3B. HPV determination from RNASeq data

RNASeq data was available for 79 tumors and a large number of reads corresponding to different HPV types were observed in RNASeq data, indicative of the presence and active transcription of viral types in the diseased tissues. Moreover, all RNASeq sequenced samples contain reads mapping to multiple HPV types. The application of highly sensitive next generation sequencing (NGS) technologies to HPV-driven cancers will likely lead to a more accurate estimation of the true prevalence of multiple HPV types in tumors. However, it is well known that low levels of cross-sample contamination tend to occur in next generation sequencing projects[12] during the library construction process [12]. A low HPV type read count in a sample with multiple HPV types is suggestive of cross sample contamination. Therefore, we filtered out low level HPV reads in order to obtain a conservative estimate of the diversity and frequency of HPV variants present in each RNASeq sample and maintain a low false positive rate. This process likely led to a moderate false negative rate in identifying multiple HPV types in each tumor. In future work, interpreting the significance of HPV infection below the detection limit of PCR but identifiable through deep sequencing will be of interest.

The following filtering procedure was used to determine HPV status:

1. Quantification of the HPV read count and identification of integration sites in each sample: the normalized value for each HPV type in a sample was calculated by dividing the each HPV type's read count by the sample's library size (i.e. the total number of sequencing reads). High confidence HPV integration sites were identified by demonstrating evidence of at least six chimeric read pairs in which the pair mate of an HPV read mapped to the human genome for any given putative integration site For each sample, the major HPV type was defined as the type with the largest number of reads in that sample. Minor types in a sample had fewer reads mapped to a given HPV type.
2. Next, we identified cases in which a minor HPV type had evidence of only 1-2 chimeric read pairs that overlapped with a high confidence integration site involving the same HPV type in another sample. Minor types with this characteristic were classified as contaminating.
3. The normalized HPV abundance values for all minor HPV types were tabulated and the highest normalized value for a contaminating minor type among all samples was identified (highlighted in red in the "Normalized RNASeq" tab in Supplementary Table 15) . Thereafter, we removed all normalized values lower than the value for the contaminating minor type with the highest normalized value.

Based on these very stringent filtering procedures, 74 (94%) of 79 tumors were assessed as HPV positive. Comparison with DNA-based typing revealed that 73 (94%) of 78 tumors with both f-HPV and RNASeq data were HPV-positive by both methods. No tumors with RNASeq data were negative by the f-HPV test. Similarly, 69 (95%) of 73 tumors with both MassArray and RNASeq data were HPV positive by both methods. One tumor was found to be negative by both assays.

## *SN4. Sequence Data Generation*

### 4A. Whole Genome Sequencing Library Construction

We followed established protocols at the Broad Institute previously described[5,6]. Libraries were sequenced using 101 bp paired-end reads. The mean coverage achieved was 35x.

### 4B. Whole Exome Sequencing Library Construction

We followed established hybrid selection protocols at the Broad Institute previously described[5,6] which are an adaptation of the procedure also described previously[7]. Exome targets were generated based on CCDS + RefSeq genes (http://www.ncbi.nlm.nih.gov/projects/CCDS/ and http://www.ncbi.nlm.nih.gov/RefSeq/), representing ~18,560 genes (93% of known, non-repetitive protein coding genes) and spanning ~1% of the genome (32.7 Mb). Libraries were sequenced using 76 bp paired-end reads. The mean coverage achieved in target regions was 89x.

### 4C. cDNA Library Construction

We followed established protocols at the Broad Institute previously described[8]. Libraries were sequenced using 76 bp paired-end reads. The mean coverage achieved was 51x.

### 4D. Sequencing

As previously described[5,6,8], we followed the standard protocol of the Broad Institute to generate initial reads and qualities directly from the sequencer. Samples on the flowcells were sequenced with Illumina HiSeq 2000 using the V3 Sequencing Kits and the Illumina 1.3.4 pipeline. All libraries were sequenced in this manner.

## *SN5. Single-Nucleotide Polymorphism (SNP) Array Based Analysis*

Genomic DNA from tumor and normal samples was processed using Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc.) according to manufacturer's protocols. Array preparation, processing, and generation of quality control (QC) metrics were performed at Genomics Platform of the Broad Institute[5,6,9].

Copy number segment edges are detected from Affymetrix SNP6 arrays using circular binary segmentation algorithm[10] as implemented in the ParallelCBS Genepattern module (ftp://ftp.broadinstitute.org/pub/genepattern/modules_public_server_doc/ParallelCBS.pdf) of the Broad Genepattern Affymetrix SNP6 Copy Number Inference Pipeline (http://www.broadinstitute.org/cancer/software/genepattern/modules/snp6copynumberpipeline). The workflow calibrates signal intensities, computes genotypes at SNP probe sites, removes outlier probes from copy number analysis, normalizes the tumor signal using a large collection of normal tissue SNP6 arrays, and finally segments the somatic copy number profile using the circular binary segmentation algorithm. Copy number segments in the algorithm must contain at least two SNP probes. Affymetrix SNP6 probes are separated by an average of 1500bp across the genome and the breakpoint resolution is limited by the local probe density at a given region.

Of the 236 samples analyzed in this project, 160 had corresponding array data (80 tumor normal pairs) and, of those samples, 153 passed SNP QC. Seventy-one tumor normal pairs with SNP 6.0 data passed both SNP and sequencing QC. Available array data was used to estimate cross-individual contamination, using ContEst[11], as described below and to analyze genome amplifications near integration sites.

## *SN6. Sequencing Data Analysis*

### 6A. BAM File Generation ("Picard Pipeline")

We used Picard (http://picard.sourceforge.net/), the standard data pipeline of the Broad Institute, to generate BAM files[12] (http://samtools.sourceforge.net/SAM1.pdf) for each tumor and normal sample. BAM files contain information on all of the reads, including the alignment to the genome, quality score, and pair orientation. BAM files were generated for Whole Exome Sequencing (WES), Whole Genome Sequencing (WGS), and cDNA sequencing results (RNASeq). RNASeq reads were aligned using TopHat[13] with hg19 as the reference genome. Detailed descriptions of this process can be found in previous publications[5,6,14,15].

### 6B. The Cancer Genome Analysis Pipeline ("Firehose")

This pipeline runs a set of tools on matched tumor-normal pairs of BAM files. There has been previous work describing the Cancer Genome Analysis Pipeline[5,6,9,14,15]. A short summary and differences from previous work, specific to this project, are described below (Supplementary Fig. 1):

1. Quality control – Determines if there has been a mix-up between tumor and normal for the same individual. Additional quality metrics for RNASeq data were generated by RNASeQC[16].

2. Tumor purity and ploidy estimation ("ABSOLUTE") – Observed copy number profiles in SNP Array data were analyzed using ABSOLUTE[17]. Results of the purity and average ploidy estimates can be found in (Supplementary Table 3).

3. Contamination Estimation ("ContEst")—Estimates for cross-individual contamination were generated for all WES and WGS using ContEst[11]. In this project, ContEst was configured in two ways: with and without SNP Array data. In the former, the standard ContEst was run. The latter configuration was run to create estimates for samples missing array data or where the array failed QC. Contamination estimates could be generated without the array data by choosing the same sites as used by the SNP 6.0 Array and genotyping the normal sample using the sequence data. By assuming germline variants as homozygous non-reference, the alternate reads represent contamination from another individual and statistics can be generated on the percentage of contamination.

4. Single nucleotide variant (SNV) calling ("MuTect") – Somatic alterations were identified by a statistical analysis of the base (and quality) at each read, the contamination estimates, and the comparison between Tumor and Normal samples. Previous work describes MuTect [5,6,14,18] which can be found at (**http://www.broadinstitute.org/cancer/cga/mutect**). Performance metrics for MuTect

revealed that the sensitivity for detecting a mutation of allelic fraction of 0.1 at a tumor read depth of 80X is 0.99.

5. Identification of small insertions and deletions ("Indelocator") – Small somatic insertion and deletion events are first identified in the tumor sample and identical events appearing in the normal sample are removed. Descriptions of Indelocator can be found in previous work[5,6,14] and at (http://www.broadinstitute.org/cancer/cga/indelocator).

6. Removal of oxidation artifact from single nucleotide variant calls ("D-ToxoG") – During the project, an artifact, resulting from oxidation during library construction, was discovered by the Broad Institute Sequencing Platform. A detailed description of the oxidation chemistry leading to this artifact is found in previous work[19]. The artifact only affects WES SNV calls. Therefore, we did not run D-ToxoG (http://www.broadinstitute.org/cancer/cga/dtoxog) on SNV calls based on whole genome or RNA sequencing. A subset of the samples in this project were affected, but each at different severity. The oxidation reaction leads to additional C>A or G>T SNV calls with low alternate read counts, with a high correlation between mutation mode and read pair orientation, and that do not validate[20]. We processed all SNV calls using D-ToxoG, and removed calls flagged as artifacts. The target rate for passing artifactual SNVs was set to one percent of the total mutations in a given tumor-normal pair. In brief, D-ToxoG consists of the following steps:

    a. For each SNV call, define $F_{oxog}$ as the number of alternate reads in the preferred artifact pair orientation (F2R1 for C>A and F1R2 for G>T) divided by the total number of alternate reads. Statistically, $F_{oxog}$ should be 0.5 at non-artifact sites, since read pair orientation is independent of the nucleotide read.

    b. Estimate the number of true artifact calls ($N_{oxo}$) in each tumor-normal pair. For each alternate allele count (ac), fit the best weight ($x_a$) to a binomial mixture model

    $$(1 - x_a)B(ac, .5) + x_a B(ac, P_{oxoG})$$

    to match the distribution of observed calls with the given alternate allele count. The first term represents non-artifact calls, hence can be modeled as a binomial distribution with equal probability of pair orientations (F1R2 or F2R1). The second term is the distribution of actual artifacts. $P_{oxoG}$ is the probability of an artifact having the corresponding pair orientation (F1R2 for G>T and F2R1 for C>A) and was set to 0.96. Before the current study, this value was determined empirically across a set of more than 500 tumor-normal pairs with high levels of artifact contamination. $P_{oxoG}$ is assumed to be constant across tumor-normal pairs, since it models the artifact generation during sequencing and all pairs in the current study went through the same process.
    $N_{oxo}$ is the sum of the artifact component (i.e. the second term of the binomial mixture model). If $N_{oxo}$ is less than one percent of the total SNV calls, then stop at this step.

c. For each C>A or G>T SNV call, calculate the p-value ($pox_m$) that the candidate call (m) is an artifact:

$$pox_m = B_{CDF}(F_{oxog}m; m_{ac}, .96)$$

Here $F_{oxog}m$ is the $F_{oxog}$ for the candidate call, $m_{ac}$ is the alternate allele count for the candidate call, and $B_{CDF}$ is the binomial cumulative density function.
For all other SNV calls, set $pox_m$ to zero.

d. Calculate the FDR[21]. Keep only those calls with a q-value less than 0.01. Therefore, the remaining cases are expected to contain no more than one percent of artifacts.
The number of SNV calls filtered by D-ToxoG can be found in Supplementary Table 3.

7. Annotation of mutations ("Oncotator") – All somatic SNV and Indel calls were annotated based on publicly available databases. Variants were mapped to genes, transcripts, and other features. These features were used to generate another set of annotations to predict a change to the protein product, if one existed. All transcripts for a given site were included, but only the GAF canonical transcript was used in downstream processing. Reference transcripts were taken from the TCGA GAF 2.1 hg19 June 2011 (https://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle), which is derived from the UCSC Genes Track[22]. Variants were also annotated with information from dbSNP build 134[23], UniProt Release 2011_09[24], COSMIC v55[25], Tumorscape[26], the Cancer Gene Census[27], the Familial Cancer Database[28], Human DNA Repair Genes[29,30], ORegAnno UCSC Track[31], MutSig Published Results[5,9,15,32,33], and the cancer cell line genotypes from the Broad-Novartis Cancer Cell Line encyclopedia (http://www.broadinstitute.org/ccle).

8. Calculation of mutation rates – We calculated aggregate mutation rates from the number of total number of SNV and indel mutations divided by the total number of bases covered.

9. Identification of significant gene mutations ("MutSig 2.0") – For each gene, we calculated the probability of seeing the observed constellation of somatic mutations or a more extreme one, given the background mutation rates calculated across the dataset. This procedure has been described previously[5,6,34]. In this project, we set a q-value threshold to 0.1, which sets the expected FDR to 10 percent[21]. No calls flagged as artifacts by D-ToxoG were used as input to MutSig nor were flagged calls used to calculate mutation rate.

10. Identification of inter- and large intra-chromosomal structural rearrangements ("dRanger") – Groups of paired-end reads which connect genomic regions with an unexpected orientation or distance are used to identify rearrangements. Details of this procedure can be found in previous work[15] and on the dRanger website (http://www.broadinstitute.org/cancer/cga/dRanger).

11. Identification of focal and arm-level regions of somatic copy number alterations ("GISTIC2.0") – In previous work[6,14], array data was used to create segment files that would be processed by GISTIC2.0. For this project, we created segment files from sequencing data in order to include all tumor-normal sample pairs. A relative copy number ratio between tumor depth and normal depth is made at each exon with sufficient coverage. Segment edges are created using circular binary segmentation[10]. Segments were used as input to GISTIC2.0, which identifies focal and broad somatic copy number alterations[26].

12. Germline calling of SNVs ("UnifiedGenotyper") – Germline mutation calling was performed as previously described[35,36] on all normal WES samples from tumor-normal pairs. Germline significance analysis was not performed due to lack of a suitable background model.

We processed whole exome and whole genome files through the entire pipeline. Since RNASeq data were being used for validation of mutations and HPV analysis, we only ran steps 1, 3, and 4 on the RNASeq data after the generating the variant calls.

## 6C. Gene Expression Analysis

Gene expression values were generated from RNA sequencing data for 17,327 genes with HUGO symbols using Cufflinks V 2.0.2[37]. The gene expression data was obtained as fragments per kilobase of exon per million fragments mapped (FPKM) values at gene level and normalized by upper quantiles. Subsequent downstream analyses were performed under $\log_2$ transformation.

It is known that mutations tend to accumulate in non-transcribed genes[9,15,38,39]. It is thus likely that genes with low expression may have a higher frequency of artifactual mutations. We therefore utilized a gene expression-based filtering as a means of increasing our confidence in the genes reported as statistically significant, as we have reported previously[40]. Based on this, genes were determined to be significantly mutated if recurrent somatic mutations were found in that gene at a false discovery rate of $q<0.1$ after correction for multiple hypothesis testing, as previously described[6,9] (Supplementary Note 6B), and if the gene had a median expression value of at least 1 fragment per kilobase of exon per million fragments mapped (FPKM $\geq$1) (Supplementary Fig. 6).

## 6D. Mutation Validation

We analyzed validation data (Targeted Resequencing and RNASeq) for a total of 85 point mutations and indels (Supplementary Table 6), which had enough reads to have confidence in the presence of the variant. For a mutation to be considered "validated", it needed to be validated in at least one of the approaches described below:

### *Approach 1: Targeted Sequencing*

Sixty two mutations were selected for targeted resequencing based on their appearance in the initial MutSig significant gene list (Supplementary Tables 7 and 9). Targeted resequencing of selected mutations for validation was performed by PCR using a microfluidic device (Fluidigm), following the manufacturer's instructions. PCR primers were designed with 200 bp flanking tails around mutations of interest. All amplicons for a given sample were given the same barcode. Constructed libraries were loaded onto an Illumina MiSeq and sequenced using paired-end 150 bp reads.

Validation was performed by manual review using visualized regions of the genome (Integrative Genomics Viewer[41]). Mutations were considered validated if supported by five or more reads.

***Approach 2: RNA Sequencing (RNASeq) and Whole Genome Sequencing (WGS)***

In order to investigate those mutations not validated by targeted resequencing and to provide extra validation, we validated mutations in the exome against RNASeq and WGS data. For every sample where RNA or WGS data was available, we compared the SNVs between the exome and available validation data (RNASeq and/or WGS). A mutation was considered "validated" if it appeared in the validation data under the following conditions:

1) There were at least two reads with the same alternate allele at the same location in the validation data. ($N_{\text{val}} \geq 2$)
2) The power (W) to call the mutation at that location was greater than 80%. The power is the probability to observe at least $N_{\text{val}}$ reads in the validation data, given the coverage at that site, if the variant is actually present.

The power is calculated by complementing the sum of the probabilities of seeing fewer reads than $N_{\text{val}}$, given coverage for the original and validation datasets and the number of alternate reads seen in the original dataset (Eq. 1)

$$W = 1 - \sum_{k=0}^{N_{\text{val}}-1} P(k \mid \tilde{n}, n, n_{\text{alt}}), \qquad (1)$$

Where $\tilde{n}$ is the coverage at the site in validation data, $n$ is the coverage at the site in the original data, $n_{\text{alt}}$ alternative-allele read count in the original data, k is the specific number of reads seen, and $N_{\text{val}}$ is the minimum number of reads required.

The probability of seeing a particular alternative-allele read count (k), can be modeled as a beta-binomial distribution. This allows calculation of the sum term (Eq. 1) and, subsequently, power. A manuscript describing the validation process and derivation of the beta-binomial distribution is being developed (Sivachenko et al. 2013 in preparation).

Corresponding RNASeq and WGS data were available for 13,346 and 1,529 somatic mutations, respectively, and our investigation of the mutation sites with adequate power for validation (Supplementary Figure 7) revealed validation frequencies of 0.85 and 0.92 by RNASeq and WGS, respectively. For mutations investigated by both RNASeq and WGS, The validation frequency by both methods was 0.83 while the validation frequency for at least one method was 0.998 (Supplementary Figure 7).

## 6E. Additional *HLA* Validation

Due to the highly variable nature of the human leukocyte antigen (*HLA*) genes[42], alignment artifacts are more common. In order to detect sequencing artifacts in the mutations called in the *HLA-A* and *HLA-B* genes, all non-silent mutations in these two genes were manually reviewed using the following procedure:

a. Identify other nearby variants and decide how many haplotypes are nearby.
b. Consider only shared reads between other variants and our candidate mutation. Candidate variant mutations should have a total correlation to the nearby variants or a total anti-correlation. In other words, the candidate variant mutation should be in its own haplotype or share a haplotype with other well-correlated nearby variants.

c. No candidate mutation should be seen only on reads going in one direction.

d. No candidate mutation should have poor base quality reads if there are neighboring, identical bases.

Of the eight non-silent *HLA-B* mutations, one frame shift insertion (patient SGCX-NOR-021, chromosome 6, position 31323362), failed the second condition. This mutation spanned 2 haplotypes, one that included a feature at position 31323012 on shared reads and one that did not. This mutation is not included in somatic mutation significance calculations and the q value for *HLA-B* significance was still < 0.1. This mutation did validate using RNA-Seq.

One nonsense mutation in HLA-A (patient SGCX-MEX-004, chromosome 6, position 29911087) failed the second condition. There was only a partial correlation with a SNP at position 29911064. This mutation was not validated and is not included in the significance calculations.

## *SN7. HPV Integration Site Determination*

The PathSeq algorithm[43] was used to perform computational subtraction of human reads, followed by alignment of residual reads to a combined database of human reference genomes and HPV reference genomes, resulting in the identification of reads mapping with high confidence to HPV genomes. Both RNASeq data and whole genome sequencing data were analyzed to identify HPV integration sites. Human reads were subtracted by first mapping reads to a database of human genomes using BWA, Megablast and Blastn[44,45]. Only sequences with perfect or near perfect matches to the human genome were removed in the subtraction process. To identify HPV reads, the resultant non-human reads were aligned to a database of multiple HPV reference genomes with Megablast and Blastn. HPV reference genomes were obtained from the Human Papilloma Virus Episteme (pave.niaid.nih.gov)[46]. Chimeric human and HPV read pairs were identified by extracting the pair mates of HPV reads and aligning the paired end reads to a combined human and HPV reference genome, using BWA[12,44]. The chimeric read pairs, in which one read maps to the human genome and the mate maps to the HPV genome, represent integration sites. We defined high confidence HPV integration sites as those demonstrating evidence of at least six chimeric read pairs in which the pair mate of an HPV read mapped to the human genome for any given putative integration site.

HPV integration has previously been shown to occur preferentially at fragile sites in the human genome[47-51]. We therefore also identified fragile sites within 5Mb of each integration site using the fragile site database on www.genatlas.org and other published literature[47-51]. The distance is calculated as the minimum distance between the integration site and the cytogenetic band of the fragile site.

## *SN8. Mutational Signature and Mutation Rate Analysis*

### 8A. Tp*C and *CpG mutational signatures

We performed hierarchical clustering of all 115 samples by nucleotide mutational context using the heatmap.2 function from the gplots library (http://cran.r-project.org/web/packages/gplots/index.html) implemented in R 2.15.1. Mutation counts were scaled within each sample (i.e. converted to fraction of mutations corresponding to each category) and clustered using Ward's minimum variance method[52]. There were 5 major clusters (Supplementary Fig. 4). Clusters I and II are similar and are characterized primarily by Tp*C to T/G mutations. Cluster III is characterized by tumors with few mutations spread across several mutational contexts. Cluster IV has a high relative frequency of *CpG to T mutations. Cluster V tumors have both (Tp*C and *CpG ) mutational signatures. These signatures are similar that those previously described in breast cancer[53,54]

We observed that Tp*C mutations were present at a relative frequency of >0.5 in 53 (46%) tumors, compared with 10% reported in breast cancer[53,54] and that the relative frequencies of Tp*C mutations in the tumors were positively correlated with mutation rates (Supplementary Fig. 5). Conversely, the relative frequencies of *CpG mutations were negatively correlated with mutation rates, especially in squamous cell carcinomas. (Fig. 1, Supplementary Fig. 5). Note that Tp*CpG mutations are subsets of both Tp*C and *CpG mutations. For display purposes in Figure 1, Tp*CpG mutations are redistributed proportionately to each group, based on the relative frequencies of the other Tp*C and*CpG mutations in each tumor.

Furthermore, there were differences in the rates of these mutational signatures by histological type (Supplementary Table 4). 5648 (54%) of the 10328 non-silent mutations observed in squamous cell carcinomas were Tp*C to T/G mutations with a rate of 18.1 mutations per Mb at sites with a Tp*C context. The predominant mutations in adenocarcinomas were *CpG to T mutations, present a rate of 8 mutations per Mb (Supplementary Table 4) at *CpG sites.

Analysis of the relationship of epidemiological data to the mutational clusters revealed an association with age (Kruskal Wallis p = 0.035), with tumors harboring predominantly Tp*C mutations (clusters I and II) being observed in older patients (median 50.3 years) compared with tumors with the mutational patterns (median 42.8, 42.2 and 45 years for clusters III, IV and IV respectively). In addition, the mutational patterns were also associated with histology (Fisher's exact p = 0.026); 13 of the 24 adenocarcinomas were enriched for *CpG mutations while 21 of the 23 tumors with the Tp*C predominant mutational pattern were squamous cell carcinomas. There was no association of mutational clusters with smoking (Fisher's exact p = 0.988), tumor grade (Fisher's exact p = 0.796), or geography (Fisher's exact p = 0.304).

Traditionally, the non-squamous carcinomas of the uterine cervix, including adenocarcinoma and adenosquamous carcinoma have been treated as a unit[55,56]. However, the major mutational type was in the Tp*C context, as opposed to the preponderance of *CpG to T mutations in adenocarcinomas. Therefore, we decided not to combine the adenocarcinomas with the adenosquamous carcinomas for the mutation significance analyses because of this difference.

### 8B. Epidemiological factors influencing mutation rates

The aggregate mutation rate of 3.7 mutations/Mb in this cervical cancer cohort is similar to rates observed in head and neck squamous cell carcinomas (3.3 mutations/Mb), but higher than the rates in pediatric rhabdoid (0.19 mutations/Mb)[14], breast (1.27 mutations/Mb)[6] and prostate (1.4 mutations/Mb)[15] cancers. It is however much less than the mutation rate in lung adenocarcinoma (12 mutations/Mb)[57]. See Supplementary Table 4.

There was a statistically significant difference between the mutation rates for squamous cell carcinoma (mean=4.2/Mb) and adenocarcinoma (mean=1.6/Mb), with a Wilcoxon p value of 0.0095 (Supplementary Table 4).

We sought to adjust the statistically significant mutation rate across histology (Wilcoxon p=0.009 Supplementary Table 4) for common epidemiological factors (Supplementary Table 5). To address this, we considered four factors across histology: patient age (age), tumor grade (grade), geographical location of sample collection (geography), and smoking status at diagnosis (smoking). Since there was an asymmetric distribution of mutation rate across the dataset, all epidemiological analyses were done using a transformed mutation rate of $\log_{10}(m + 1)$, where m is the actual nonsilent mutation rate in units of mutations/megabase of sequenced DNA. We found that the squamous cell carcinomas were from an older population (Wilcoxon p = 0.042, median age was 47.8 years versus median age of 41.0 years in adenocarcinoma patients). Due to a low number of tumor grade 1 samples and a concentration of grade 1 in the adenocarcinomas samples (5/6 of grade 1 samples were adenocarcinomas), we did observe an association between tumor grade and histology (Fisher Exact Test p=0.002). When we applied the same statistical test again but without the grade 1 tumors, we found that the correlation between grade and histology was no longer significant (Fisher Exact Test p = 0.590). For all other tests, we included the grade 1 samples. We were unable to confirm an association between smoking status and histology (Fisher Exact Test p= 0.12) or mutation rate (Wilcoxon p = 0.467), though this may have been due to lack of statistical power and/or inconsistent reporting of exposure across subjects. Geography was found to be a factor for the transformed nonsilent mutation rate (Wilcoxon p = 0.037 Norway: 0.41269 Mexico: 0.60131), but this may be due to the higher median age of the Mexican cohort (Kruskal-Wallis p = 0.017; Median age: Norway = 43.5, Mexico = 52). Similar to results in other work[53,58], we observed a correlation between age and mutation rate in our cohort (Pearson correlation p = 0.005, $R^2 = 0.08$). Since we only saw associations between mutation rate, histology, and age, we used a linear regression model (LRM) to test whether histology and age are independently significant predictors of mutation rate (Equation 2).

$$\log_{10}(m+1) = c + B_0 * x_1 + B_1 * x_2 \qquad (2)$$

where m is the actual nonsilent mutation rate, c is a constant term, $x_1$ is the histology (adenocarcinoma = 1, squamous cell carcinoma = 0), $x_2$ is age (continuous), $B_0$ is the fitted coefficient for the histology term and $B_1$ is the fitted coefficient for the age term.
On testing for histology (adenocarcinoma vs. squamous cell carcinoma) and age as predictors of mutation rate, both were significant (histology p = 0.045 and age p = 0.012). Age was a positive correlate for mutation rate ($B_1 = 0.006$), while adenocarcinoma histology was a negative correlate ($B_0 = -0.145$). The final linear model is found in Equation 3.

$$\log_{10}(m+1) = 0.259 - 0.145 * x_1 + 0.006 * x_2 \qquad (3)$$

where m is the actual nonsilent mutation rate, $x_1$ is the histology (adeno = 1, squamous = 0), and $x_2$ is age (continuous).

## *SN9. Significantly Mutated Gene Sets*

This is, to our knowledge, the largest set of cervical carcinomas investigated by exome sequencing to date. Nonetheless, the relatively small sample sizes probably limited our ability to identify other significantly mutated genes in our dataset. We have therefore included Supplementary Tables 7 and 8 to highlight other genes with false discovery rates just below the significance threshold of q<0.1. Some mutated genes of interest in this category include *TP53, CASP8, HLA-A, RB1* and *B2M* in squamous cell carcinomas, and *KRAS* and *PIK3CA* in adenocarcinomas.

Pathway analyses in the squamous cell carcinomas revealed that many significantly mutated genesets were driven by the significantly mutated genes (Supplementary Tables 10a). However, when these genes were excluded from the analysis of squamous cell carcinomas, the interferon signaling pathway was significant (q=2.88 x$10^{-5}$). Specific mutated genes include *IFNG*(1), *IFNGR1*(4), *IKBKB*(1), *JAK2*(4), *NFKBIA*(1), *RB1*(5), *TNFRSF1A*(1), *TP53*(4), *WT1*(2), as well as *JAK1*(1) and *STAT1*(1). In addition, a cytochrome-related metabolic pathway including mutations in *CYP2A13*(2), *CYP2A6*(2), *CYP2A7*(4), *NAT2*(1), *XDH*(4), was also significantly mutated.

Pathway analysis performed on the adenocarcinomas revealed the PIK3CA/PTEN pathway as significantly mutated (q = 0.0143). Mutated genes in this pathway include *AKT1*(1), *PIK3CA*(4), *PTEN*(2) and *SOS1*(1). See Supplementary Table 10b.

## *SN10. Copy Number Analyses*

### **10A. Broad and focal level copy number alterations in cervical carcinoma**

GISTIC2.0 analysis[59] revealed there were 9 broad level gains and 11 broad level losses among 79 squamous cell carcinomas (Supplementary Table 11), while there were 4 broad level gains and 8 broad level losses among 24 adenocarcinomas (Supplementary Table 12), based on a false discovery rate of q<0.25.

Genomic events common to both histological subsets include gains in chromosome arms 1q, 3q, 19q and 20q as well as losses across chromosomes 4 and 11. Squamous cell carcinomas also had significant gains of chromosome arms 1p, 5q, 8q, 14q and 20p along with losses in chromosome arms 8p, 13q and 17p as well as across chromosomes 3 and 6. Losses in chromosomes arms 18q and 19p, as well as across chromosome 16, were unique to the adenocarcinomas. Most of the SCNAs observed have been previously reported, and Supplementary Tables 11-12 includes a representative (but not exhaustive) list of relevant references.

Supplementary Fig. 10 shows a comparison of these data by histological type. Squamous cell carcinomas had higher frequencies of chromosomal arm 3p loss and 1p gain while there were higher frequencies of chromosomal arm deletions in 18q and 16q in adenocarcinomas.

Sixteen significant focal amplifications and 25 significant focal deletions were found in the squamous cell carcinomas (Supplementary Figs 11-12). Most of these events (10 /16 focal amplifications and 16/25 focal deletions) did not overlap with broad-level copy number changes

(Supplementary Figs 11-12). There were 8 significant focal amplifications of which only 3 overlapped with broad level gains in adenocarcinomas (Supplementary Fig. 13). There were no significant focal deletions in the set of 24 adenocarcinomas. The most significant focal amplification peak in squamous cell carcinomas was in chromosomal band 11q22, with 6 genes including *BIRC3* and *YAP1* (Supplementary Fig. 11). In addition, both histological types had significant focal amplifications in chromosomal bands 17q12 (a wide peak including *ERBB2*), and 8q24 (*MYC*). None of these peaks overlapped with a region of broad copy number gain. In addition, a peak in chromosomal band 1q21.3 encompassing MCL1 was focally amplified in the cervical adenocarcinomas, although it overlapped with a broad amplification peak.

Some focal peaks harbor some genes implicated by other modalities in this study. *NFE2L2* is significantly mutated (Table 1) but also lies within the 2q24 amplification peak. In addition, *EIF2C2* and *GLI2* are both associated with HPV integration (Fig. 3), and are part of the 8q24 and 2q14 amplification peaks respectively.

## 10B. Hierarchical Clustering of copy number data

Hierarchical clustering of the copy number profiles of all tumors revealed 3 significantly different subsets (high, intermediate, low) based on the relative frequency of altered copy number segments (Supplementary Figs 15-16). We did not find any association between the copy number clusters and histology (Fisher's exact p=0.7125) or tumor grade (Fisher's exact p = 0.6686).

## 10C. Genome-Wide Correlation Analyses of Copy Number Alterations and Gene Expression

We performed a "genome-wide" Pearson correlation analysis between DNA copy number changes and RNASeq-derived gene expression (FPKM) for all tumors with both types of data available, and calculated false discovery rates[21,60] for each correlation. These data are presented in the attached Supplementary Table 13. Examples of these relationships are also presented in Supplementary Fig. 17. Only 6874 (41%) of 16898 genes investigated had significant correlative relationships at a false discovery rate of $q<0.05$, with the lowest corresponding correlation coefficient (r) values being +0.22 and -0.22. Some of the significantly mutated or altered genes in this study such as *ERBB2, KRAS, PIK3CA, PTEN, FBXW7, STK11* and *BIRC3* had r values greater than 0.22. Other genes including *TP53, MYC, EP300, RB1* and *MAPK1* had $r<0.22$, with q values above 0.1 (Supplementary Fig. 17). In particular TP53 showed no correlation between copy number and gene expression with $r=-0.03$ and $q=0.36$ (Supplementary Fig. 17). The gene with the highest positive correlation was *MIEN1*, with $r=0.74$ and $q=2.77 \times 10^{-11}$. This gene is immediately downstream of *ERBB2* on chromosomal band 17q12. On the other end of the spectrum, several genes involved in the immune response (exemplified by *SHC1* and *CD2*) as well as within chromosomal band 1p36 (exemplified by *RUNX3*) had the most negative correlation between copy number and gene expression (Supplementary Table 13). We also explored these relationships in the context of HPV integration sites (See Supplementary Note 11G and Supplementary Figs 25-28).

## 10D. Relationship between Copy Number changes and Somatic Mutations

We examined the relationship between copy number alterations and somatic nonsilent mutations for a subset of significantly mutated genes (Supplementary Fig. 14). For some genes, there was some overlap between copy number changes and somatic mutations. For example, 4 of

5 patients with *ERBB2* mutations have low level copy number gains while 3 of 4 patients with *STK11* mutations have low level copy number loss. All tumors with *HLA-B* mutations had some level of copy number loss, while no *EP300* mutations occurred in samples with decreased *EP300* copy number. Furthermore 3 of the 4 tumors with *NFE2L2* mutations and 2 of the 3 tumors with *ELF3* mutations had copy number gains (Supplementary Fig. 14).

## *SN11. HPV Integration Events*

### 11A. General HPV Integration Patterns

We observed 65 HPV integration sites, including several within or in close proximity to fragile sites, as well as previously reported genes[47,48,50,61]. There were a few hotspots for HPV integration. For example, we observed HPV integration events in chromosome cytoband 8q24 involving several genes including *MYC*, *EIF2C2*, *MAFA* and a lincRNA (*LOC727677*). Other chromosomal loci with recurrent HPV integration events include 1q32 (*PROX1)*, 2q22 (*FRA2K*), 3q25 (*RAP2B, FRAD3D*), 9q22 (*FANCC, C9orf3, LINC00475*), 9q34 (*EGFL7, LOC100506190*), 14q24 (*RAD51B*), 17q12 (*ERBB2*), 17q21 (*RARA, KRT39*) and 19q13 (*CEACAM5*).

The locations and frequency of these events are reported in Supplementary Table 15.

### 11B. Recurrent HPV integration into the *RAD51B* gene

We observed recurrent HPV integration in the intronic regions of the *RAD51B* gene in three different tumors, each involving a different HPV type: HPV16, HPV18, and HPV52 (Supplementary Table 15). *RAD51B* is located in the same chromosomal cytoband, 14q24.1, as the fragile site *FRA14C (*Supplementary Table 15). While an HPV integration event involving *RAD51B* has been reported[48], the recurrence involving 3 different HPV types was striking. See Circos plots[62] and HPV-human fusion transcripts in Supplementary Fig. 19-20. *RAD51B* is required for DNA repair by homologous recombination[63] and variants of this gene have been associated with susceptibility to cancer[64]. It is therefore conceivable that disruption of this gene by viral integration may facilitate tumor development.

We designed primers based on the two halves of each set of HPV-*RAD51B* chimeric reads in the 3 respective tumors, and successfully amplified 2 of the 3 fusion transcripts (Supplementary Fig. 23) from tumor cDNA (derived from a fresh aliquot of the original RNA). All 3 tumors and an additional tumor without HPV-*RAD51B* integration were investigated by each set of primers. The amplification was specific to only the tumor with the appropriate chimeric read. The tumor for which RT-PCR validation failed, had only 9 chimeric read pairs, compared with the other 2 tumors with 145 and 245 chimeric read pairs respectively. This result may therefore be due to the limit-of-detection of low abundance reads by PCR.

### 11C. HPV Integration into genomic fragile sites

We observed that 34 of the 65 integration occurred at locations up to 5Mb away from fragile sites[47-51]. The fragile sites involved include *FRA1G, FRA1I, FRA2E, FRA2K, FRA3B (FHIT), FRA3D, FRA3C, FRA4F, FRA6F, FRA7D, FRA7I, FRA8D, FRA9B, FRA9D, FRA12B, FRA13E, FRA14C, FRA16A, FRA17B, FRA18B, FRA19A, FRA20B, FRA22A* and *FRAXB*. These data are presented in Supplementary Table 15, and are similar to previous reports of HPV integration occurring within or in close proximity to fragile sites in human genomes[47,48,50,61]

## 11D. Co-occurrence of HPV Integration and Copy Number Amplifications

We observed that a large fraction of integration sites were supported by more than 20 spanning read pairs (Supplementary Table 15) and, furthermore, that many of the integration sites occurred at locations with copy number amplifications. We hypothesized that the distance between integration sites and copy number amplifications in the genome was much shorter than would be expected if these events were each distributed randomly. Many integration sites occurred outside of exonic regions, thus precluding the use of copy number status as determined by read depth analysis. Therefore our analysis was limited to the subset of 51 tumors with chimeric human-HPV read pairs and SNP array copy number data.

The following procedure was employed:

1. We calculated the distance between each integration site and the nearest amplification. For the purposes of this test, we considered segments with a log segmean value over 0.5 to be amplified.
2. We performed a permutation test in order to evaluate our hypothesis. The locations of all integration sites were permuted uniformly across the genome while preserving the lengths of the true integration sites. These simulated integration sites were then randomly assigned to samples. In addition, the observed location and size of each amplification event were utilized in the permutation analysis.
3. We calculated the distance between each integration site and the nearest amplified region in the sample to which the integration site was assigned. In both the true data and the permuted data, if an integration site occurred on a chromosome without an amplification, we assigned this integration site a distance equal to the length of chromosome 1 in hg 19 plus 1 (1 more base than the longest possible distance between an amplification and an integration site). If an integration site overlapped with an amplified region, we assigned this integration site a distance of 0.
4. We performed 100,000 such permutations, and the true and simulated distances are plotted in Fig. 3a. The Mann-Whitney U test was utilized to evaluate whether the true distances are shorter than the distances we would have expected to observe by chance in the permuted data.

The true distances between integration and amplification were significantly shorter than the permuted distances ($p < 2.2\times10^{-16}$).

## 11E. HPV Integration and Expression of Integration Site Genes

We observed that, in tumors with HPV integration, if the integration involved a human gene, the gene was very highly expressed. For the genes with HPV-human chimeric read pairs, that gene was frequently an outlier in expression (as evaluated by log2 FPKM from RNASeq data) and these results are shown in Fig. 3b. In order to test whether genes with HPV-human chimeric read pairs tended to have high expression in the tumor with integration as compared to all other tumors, we performed a permutation test as follows:

1. We ranked the tumors in order of their $\log_2$(FPKM) for each gene with a high-confidence integration site in any tumor. Rank 79 corresponds to the highest expression and rank 1 corresponds to the lowest expression.
2. We identified the ranks of the samples with HPV-human chimeric read pairs in the 41 human genes reported in Fig. 3b (Supplementary Fig. 23 and 24).

3. To test whether these ranks were higher than would be expected by chance, we repeatedly sampled 41 ranks from the expression rank matrix without replacement and compared these two distributions of ranks. We performed 10,000 such samplings. If genes with HPV-human chimeric read pairs were randomly distributed with respect to rank, we would expect to see a uniform distribution of ranks (as in the sampled data). The two distributions of ranks are displayed in Supplementary Fig. 22.
4. In order to test whether the true ranks were greater than the ranks that we would expect by chance, we used a Mann-Whitney U test comparing the true and sampled ranks, and the result was highly significant ($p > 2.2 \times 10^{-16}$).

This results suggests, in general, that genes involved in HPV integration events have higher expression in tumors with HPV integration than in tumors without HPV integration in that gene.

## 11F. Effect of HPV Integration and Local Gene Expression

As reported in the main text and in Supplementary Note 11E, we observed that gene expression levels at sites of HPV integration were significantly higher in tumors with HPV integration compared with the expression levels of the same genes across the other tumors without integration at that site (Fig. 3b; Supplementary Figs. 22-24).

The relationship between HPV integration and gene expression was even more striking in genes with low or no expression. For example, most of the tumors have a $\log_2$FPKM value < 0 for *MAFA*, whereas the tumor with HPV integration at the *MAFA* gene has an outlier $\log_2$RPKM value of >4.

There were some differences in the integration site locations within the affected gene and the gene expression patterns of their genomic neighbors. We observed evidence of exonic chimeric read pairs in several genes including *MYC*, *ERBB2*, *GLI2*, *TNIK*, *PARN*, *EIF2C2*, *RPS6KB1*, *EGFL7*,*MAFA*, *FAM179B*, *FMO3*, *BCL11B*, *RARA*, *TP63*, *PTHLH*, *PROX1*, etc (Supplementary Fig. 23). We found no evidence of exonic chimeric read pairs for *PTPRT*, *MACROD2*, *RAD51B*, *FANCC*, *DACH1*, *KRT39*, *RAP2B*, *BCL2L13* and *C9orf44* (Supplementary Fig. 24), suggesting that the integration events occurred within introns in these genes. Interestingly most of the genes with intronic integration sites also have median-to-low relative expression levels within their respective distributions.

We observed a general pattern in which the neighboring genes, on the immediate 5' and 3' flanks of the integration site genes, had gene expression levels that were either around the median or below the median of their respective distributions (Supplementary Figs 23-24). This contrasts with the relatively high expression of the integration site genes themselves, and suggests that the effect of HPV integration on gene expression may be focal and somewhat limited to the index gene and not the neighboring genes.

Note that in some instances, we observed chimeric human-viral read pairs involving multiple contiguous genes within the same genomic locus. For example, *ERBB2* is flanked by *STARD3*, *PGAP3* and *IKZF3*, genes which also had chimeric human-HPV read pairs (Supplementary Figs 23-24).

## 11G. Relationship of Copy Number Alteration and Gene expression in the context of HPV integration

As reported in Supplementary Note 10C, we also compared somatic copy number alterations and RNASeq-derived gene expression for all tumors. The relationships between these two parameters were further analyzed with a focus on integration site genes and are reported in Supplementary Figs 25-27. We observed 3 approximate patterns of relationships:

1. Integration sites with outlier high gene expression associated with copy number gain: *MYC, ERBB2, GLI2, TNIK, EIF2C2, SERPINB4, NR4A2, PROX1, FAM179B, KRT39, P4HB, CEACAM5, GRB7, IKZF3, SERPINB3, RAP2B* (Supplementary Fig. 25)
2. Integration sites with outlier high gene expression with no copy number gain: *PARN, RPS6KB1, EGFL7, MAFA, FMO3, SNIP1, PTHLH, POC1B, DACH1, KLH28, BLVRA, TMCC3, BCL11B, PLA2G10, STARD3, PGAP3*(Supplementary Fig. 26)
3. Integration sites with miscellaneous relationships between gene expression and copy number: *TP63, PTPRT, MACROD2, RAD51B, FANCC, RARA,BCL2L13, C9orf3* (Supplementary Fig. 27)

These data highlight significant associations among HPV integration, gene amplification and gene expression across several genes, beyond the reported associations with *MYC*. This suggests that variable mechanisms may be at play in HPV-driven human gene expression including the role of genomic amplification and the viral promoter[48].

### 11H. HPV Integration Site Genes and Genesets

We interrogated the Molecular Signatures Database (MSigDB) to test if the genes involved in HPV integration events co-occur in the same biological pathways or genesets. We observed that many of these genes overlap with pathways important in cancer. See Supplementary Table 16 for details.

## *SN12. Consensus Clustering and Histology-based Gene Expression Analysis*

We performed consensus clustering on the RNASeq-derived gene expression data from 79 tumors. Gene expression ($\log_2$(FPKM)) variability was assessed in terms of the median absolute deviation across patients. The 5000 genes with the largest deviation were selected for clustering. ConsensusClusterPlus[65] analysis was performed using 1000 resampling iterations and a maximum of 25 clusters. A k of 8 was chosen due to a clear bimodal behavior associated with little change in the area under the empirical cumulative distribution, upon further increases in k (Supplementary Figure 28).

The 2 robust clusters identified by consensus clustering of RNASeq data across the 79 tumors were segregated essentially by the histological diagnoses of squamous cell carcinomas and adenocarcinomas, with a few exceptions (Supplementary Fig. 28). Comparative marker selection analysis[66] of the two main clusters (Supplementary Fig. 29) yielded very similar results to the gene set enrichment analyses[67] (GSEA) comparing the two histological types. It showed that there were similarities in the genes upregulated in cervical squamous cell carcinomas and basal cell carcinomas of the breast, while adenocarcinomas were more analogous to luminal breast cancers (Supplementary Table 17)

## *SN13. Genomic Rearrangements*

### 13A. Whole Genome Sequencing analysis

The spectrum of genomic rearrangements was investigated by analyses of whole genome sequencing data as previously described[22,45] from 14 tumor-normal pairs (Supplementary Table 18). Most tumor genomes had relatively few somatic rearrangements (median=12) and these were predominantly local intrachromosomal events (See Circos plots in Supplementary Fig. 30. Although no recurrent somatic rearrangements were found, there were several events with potential significance in cancer, including rearrangements in *NTRK2* (exons 12-14), *ARHGEF3* (exon 4) and *NIPBL* (exons 33-42) (Supplementary Table 19).

There have been previous reports of cervical cancer genomes with complex rearrangements, mostly in the context of cell lines[68,69]. It is certainly possibly that a higher frequency of complexly rearranged cervical cancer genomes will be discovered with larger sample sizes. The tumors were largely chosen based on DNA quality and availability. The histological types were well represented: 8 squamous cell carcinomas, 3 adenocarcinomas, 2 adenosquamous carcinomas and 1 clear cell carcinoma (Supplementary Table 18). In the same vein, there was an even distribution of moderately differentiated and poorly differentiated tumors. There were no obvious patterns of association of rearrangements with histology or tumor grade.

### 13B. Fusions in Transcriptome Data

We utilized the RNASeq fusion algorithm implemented in Firehose to investigate the presence of fusions in transcriptome data from 79 tumors. Putative fusion events were determined by identifying inter-chromosomal chimeric read pairs or exon-exon read pairs separated by at least 1 Mb, with pairmates in the appropriate coding strand orientation. The algorithm also identifies unmapped reads spanning the putative fusion junction. In addition, a junction reference is built by considering every possible junction in coding-coding orientation, adjusting automatically for read length to ensure that reads overlap with junctions. High confidence fusion events are defined as having at least 3 reads mapped to a junction fusion.

There were no major recurrent gene fusions in the transcriptome data. The list of fusions is reported in Supplementary Table 20. Tumor SGCX-NOR-030 had overlapping data between transcriptome and whole genome sequencing data. This tumor had an interchromosomal transcript fusion event involving *IL20RB* in chromosome 3 and *RASSF8* in chromosome 12. In addition, a tandem duplication event resulted in *MYO15B-NUP85* fusions involving genes on chromosome 17. There were several instances of intrachromosomal fusion reads between two homologous genes *FRG1* (chromosome 4) and *FRG1B* (chromosome 20), as well as between different *HLA* genes on chromosome 6. We believe these are likely alignment artifacts due to high homology between these genes.

## *SN14. Miscellaneous Genes and Pathways of Interest*

### 14A. APOBEC family

Members of the *APOBEC* family of enzymes are known to deaminate C residues preceded by a T[70]. This mutational context (Tp*C) was predominantly altered in our cohort. We therefore sought to investigate the characteristics of the *APOBEC* family in the cervical tumors we studied. There were 6 missense APOBEC family gene mutations in 4 tumors (Supplementary Table 21). Over 70% of mutations in each tumor occurred in the Tp*C context. Interestingly, these tumors include the 2 samples with the highest mutation rates in whole cohort. Strikingly, the sample with the highest mutation rate (~40/Mb) has essentially the same mutation in two different *APOBEC3* family members (S93F in *APOBEC3B* and S92F in *APOBEC3F*). We did not find any correlation between Tp*C mutation rates and the gene expression of members of the *APOBEC* family.

### 14B. Fanconi Anemia Pathway

Mutations in the Fanconi anemia (FA) pathway have been associated with E7 protein accumulation, which indicates that FA genes play a role in suppressing HPV infection[71]. It is interesting to ask whether mutations in these genes are present in our study. In our analysis of somatic mutations, we found non-silent somatic mutations in *BRIP1* (1), *FANCA* (3), *FANCD1/BRCA2* (3), *FANCE* (1), *FANCI* (3), *FANCM* (2), and *PALB2* (1). However, none of these genes were statistically significantly mutated.

We also performed germline calling on the corresponding 115 normal samples. However, we were unable to analyze the statistical significance of the germline calls, due to lack of a suitable background model. Therefore, we list germline calls in the FA genes in Supplementary Table 22A, which lists individuals with both somatic and germline mutations in the FA genes. Supplementary Table 22B also shows the number of mutations and the patients where both somatic and germline mutations occurred in FA genes

### 14C. Human Telomerase RNA Component (TERC)

*TERC* has been identified as a gene marker for genomic instability and copy number gain of *TERC* has been associated with cervical cancer[72]. *TERC* was included in two significant focal amplifications within chromosome arm 3q in squamous cell carcinomas ($q = 0.102$, 284 genes) and in adenocarcinomas ($q = 0.005$, 84 genes). We found no germline mutations in *TERC* in any of 115 cases and one somatic mutation (SNP 3:169482689, G>C) in patient SGCX-MEX-008.

## *SN15. References*

1. Cancer in Norway 2010 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2012. in *Cancer Registry of Norway* (Cancer Registry of Norway, Oslo, 2012).
2. Canadas, M.P. *et al.* Comparison of the f-HPV typing and Hybrid Capture II(R) assays for detection of high-risk HPV genotypes in cervical samples. *J Virol Methods* **183**, 14-8 (2012).
3. Yang, H. *et al.* Sensitive detection of human papillomavirus in cervical, head/neck, and schistosomiasis-associated bladder malignancies. *Proc Natl Acad Sci U S A* **102**, 7683-8 (2005).
4. Walline, H.M. *et al.* High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and, oral cavity cancers: Comparison of multiple methods. *JAMA-Otolaryngology* **In press**(2013).
5. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-60 (2011).
6. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-9 (2012).
7. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-9 (2009).
8. Levin, J.Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* **10**, R115 (2009).
9. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-72 (2011).
10. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
11. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2 (2011).
12. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
13. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
14. Lee, R.S. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J Clin Invest* **122**, 2983-8 (2012).
15. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
16. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530-2 (2012).
17. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-21 (2012).
18. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).

19. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* (2013).
20. Pugh, T.J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* **45**, 279-84 (2013).
21. Benjamini, Y.H., Yosef Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
22. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-82 (2011).
23. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
24. UniProt, C. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39**, D214-9 (2011).
25. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-50 (2011).
26. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007-12 (2007).
27. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
28. Sijmons, R.H. & Burger, G.T. Familial cancer database: a clinical aide-memoire. *Fam Cancer* **1**, 51-5 (2001).
29. Wood, R.D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291**, 1284-9 (2001).
30. Wood, R.D., Mitchell, M. & Lindahl, T. Human DNA repair genes, 2005. *Mutat Res* **577**, 275-83 (2005).
31. Griffith, O.L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* **36**, D107-13 (2008).
32. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
33. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).
34. Lohr, J.G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A* **109**, 3879-84 (2012).
35. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
36. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
37. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* (2012).
38. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-8 (2011).
39. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6 (2010).
40. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).

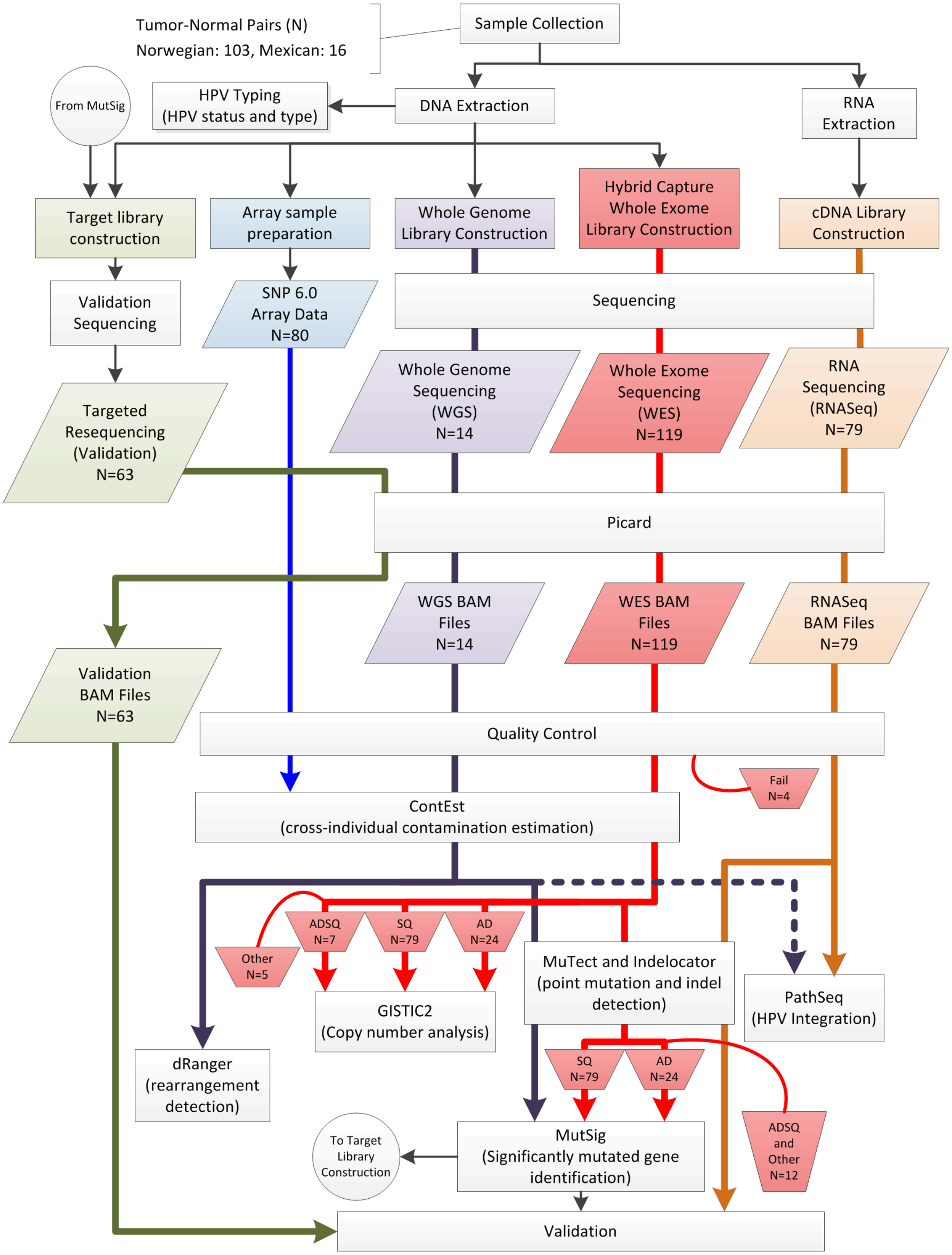
41. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* (2012).
42. Erlich, H. HLA DNA typing: past, present, and future. *Tissue Antigens* **80**, 1-11 (2012).
43. Kostic, A.D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**, 393-6 (2011).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
45. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
46. Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res* **41**, D571-8 (2013).
47. Durst, M., Croce, C.M., Gissmann, L., Schwarz, E. & Huebner, K. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. *Proc Natl Acad Sci U S A* **84**, 1070-4 (1987).
48. Kraus, I. *et al.* The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. *Cancer Res* **68**, 2514-22 (2008).
49. Peter, M. *et al.* Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J Pathol* **221**, 320-30 (2010).
50. Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I.B. & Durst, M. Non-random integration of the HPV genome in cervical cancer. *PLoS One* **7**, e39632 (2012).
51. Debacker, K. & Kooy, R.F. Fragile sites and human disease. *Hum Mol Genet* **16 Spec No. 2**, R150-8 (2007).
52. Ward, J.H., Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**, 236-244 (1963).
53. Stephens, P.J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-4 (2012).
54. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
55. Gallup, D.G., Stock, R.J. & Talledo, O.E. Current management of non-squamous carcinoma of the cervix. *Oncology (Williston Park)* **3**, 95-102; discussion 104, 106 (1989).
56. Thigpen, J.T., Blessing, J.A., Fowler, W.C., Jr. & Hatch, K. Phase II trials of cisplatin and piperazinedione as single agents in the treatment of advanced or recurrent non-squamous cell carcinoma of the cervix: a Gynecologic Oncology Group Study. *Cancer Treat Rep* **70**, 1097-100 (1986).
57. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-20 (2012).
58. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci U S A* **110**, 1999-2004 (2013).
59. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
60. Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B (Methodological)* **64**, 479–498. (2002).

61. Wentzensen, N., Vinokurova, S. & von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res* **64**, 3878-84 (2004).
62. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).
63. Takata, M. *et al.* The Rad51 paralog Rad51B promotes homologous recombinational repair. *Mol Cell Biol* **20**, 6476-82 (2000).
64. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**, 579-84 (2009).
65. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-3 (2010).
66. Gould, J., Getz, G., Monti, S., Reich, M. & Mesirov, J.P. Comparative gene marker selection suite. *Bioinformatics* **22**, 1924-5 (2006).
67. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
68. Gallego, M.I., Schoenmakers, E.F., Van de Ven, W.J. & Lazo, P.A. Complex genomic rearrangement within the 12q15 multiple aberration region induced by integrated human papillomavirus 18 in a cervical carcinoma cell line. *Mol Carcinog* **19**, 114-21 (1997).
69. Harris, C.P. *et al.* Comprehensive molecular cytogenetic characterization of cervical cancer cell lines. *Genes Chromosomes Cancer* **36**, 233-41 (2003).
70. Harris, R.S. & Liddament, M.T. Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* **4**, 868-77 (2004).
71. Hoskins, E.E. *et al.* The fanconi anemia pathway limits human papillomavirus replication. *J Virol* **86**, 8131-8 (2012).
72. Heselmeyer-Haddad, K. *et al.* Detection of genomic amplification of the human telomerase gene (TERC) in cytologic specimens as a genetic test for the diagnosis of cervical dysplasia. *Am J Pathol* **163**, 1405-16 (2003).

## Landscape of Genomic Alterations in Cervical Carcinomas

*Ojesina AI, Lichtenstein L, et al.*

## SUPPLEMENTARY FIGURES AND LEGENDS

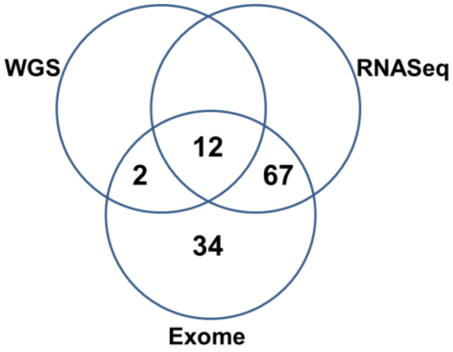

Tumor-Normal Pairs (N)
Norwegian: 103, Mexican: 16
Sample Collection
From MutSig
HPV Typing
(HPV status and type)
DNA Extraction
RNA Extraction
Target library construction
Array sample preparation
Whole Genome Library Construction
Hybrid Capture Whole Exome Library Construction
cDNA Library Construction
Validation Sequencing
SNP 6.0 Array Data N=80
Sequencing
Targeted Resequencing (Validation) N=63
Whole Genome Sequencing (WGS) N=14
Whole Exome Sequencing (WES) N=119
RNA Sequencing (RNASeq) N=79
Picard
WGS BAM Files N=14
WES BAM Files N=119
RNASeq BAM Files N=79
Validation BAM Files N=63
Quality Control
Fail N=4
ContEst
(cross-individual contamination estimation)
ADSQ N=7
SQ N=79
AD N=24
Other N=5
GISTIC2
(Copy number analysis)
MuTect and Indelocator
(point mutation and indel detection)
PathSeq
(HPV Integration)
dRanger
(rearrangement detection)
SQ N=79
AD N=24
ADSQ and Other N=12
To Target Library Construction
MutSig
(Significantly mutated gene identification)
Validation

**Supplementary Figure 1. Schematic of sample processing and analysis pipeline**
Whole exome sequencing data was generated for 115/119 tumor-normal sample pairs that passed quality control. All whole genome and RNA sequencing, SNP 6.0 Array, and validation data were produced using a subset of the remaining 115 pairs with WES data (Supplementary Figure 2). Targeted resequencing was based on the significant genes from an initial MutSig analyses. For all steps in the above schematic, N refers to the number of complete tumor-normal pairs. AD, SQ, and ADSQ indicate groupings of pairs by histology (for adenocarcinoma, squamous, and Adeno-squamous, respectively).

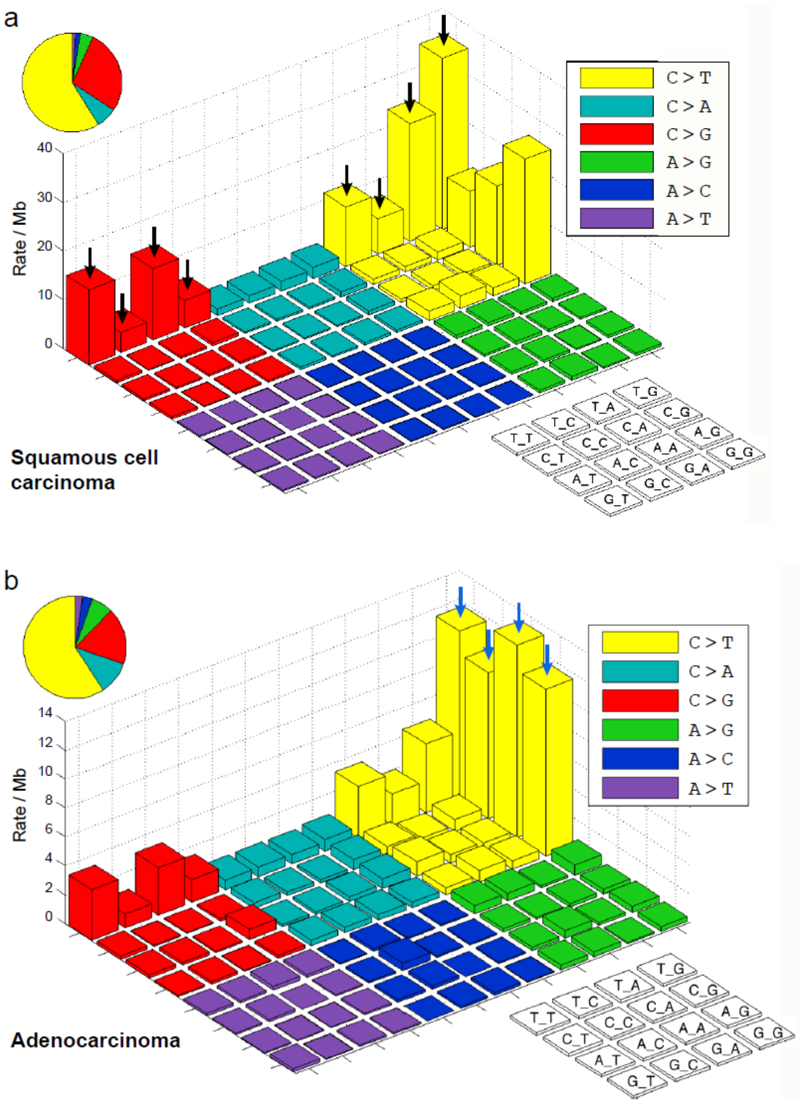
**Samples:**

100 tumor-normal pairs from Norway
100 pairs with exome data
79 samples with RNASeq data
14 pairs with WGS data

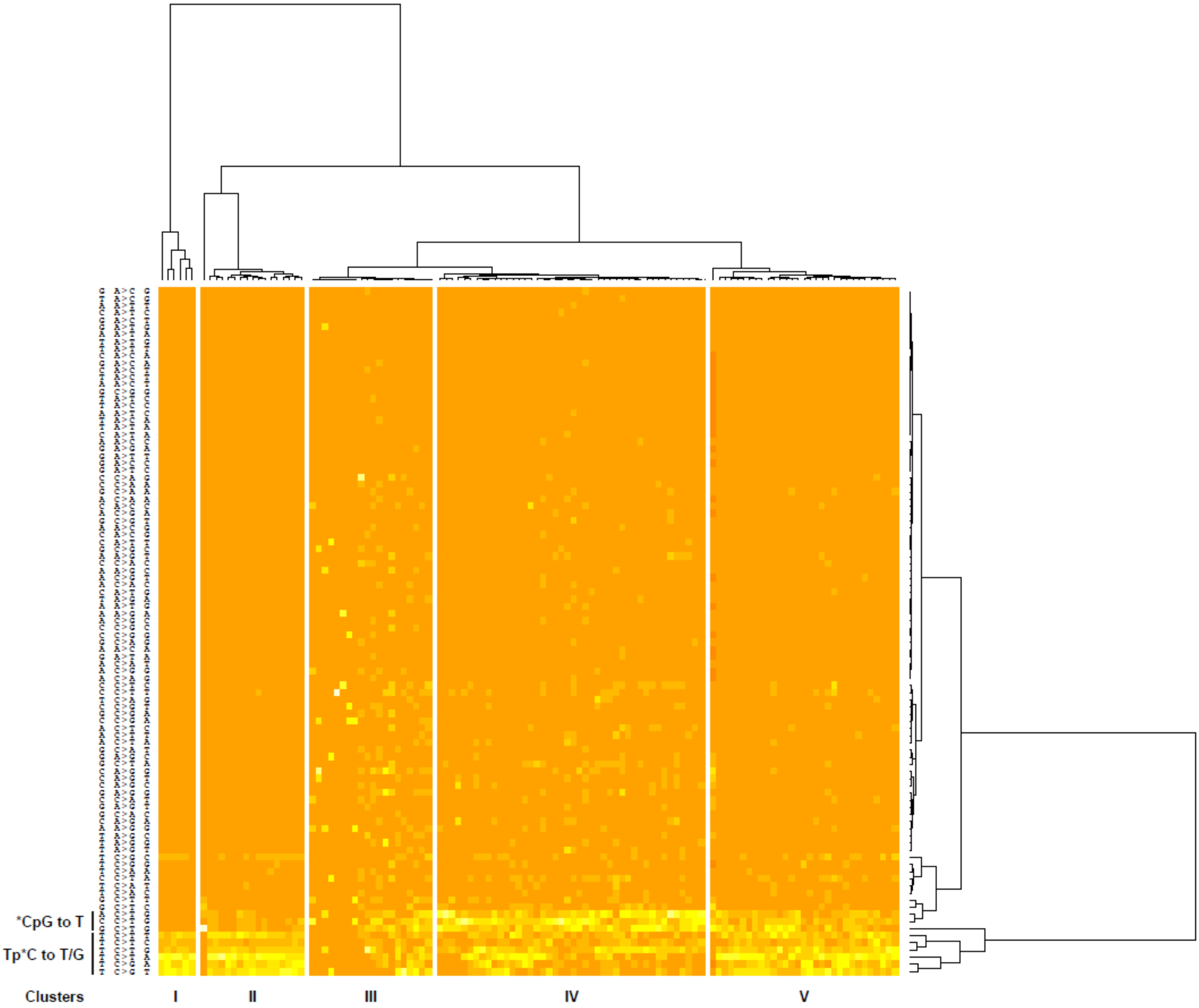
15 tumor-normal pairs from Mexico
15 pairs with exome data



**Supplementary Figure 2. Sample cohort and sequencing procedures performed**
A total of 100 primary cervical cancers from Norway and 15 from Mexico were subjected to exome sequencing (Exome). A subset of the Norwegian tumors was investigated by RNA sequencing (RNASeq; 79 tumors) and whole genome sequencing (WGS; 14 tumors). The Venn diagram shows the frequencies of tumors investigated by overlapping procedures.

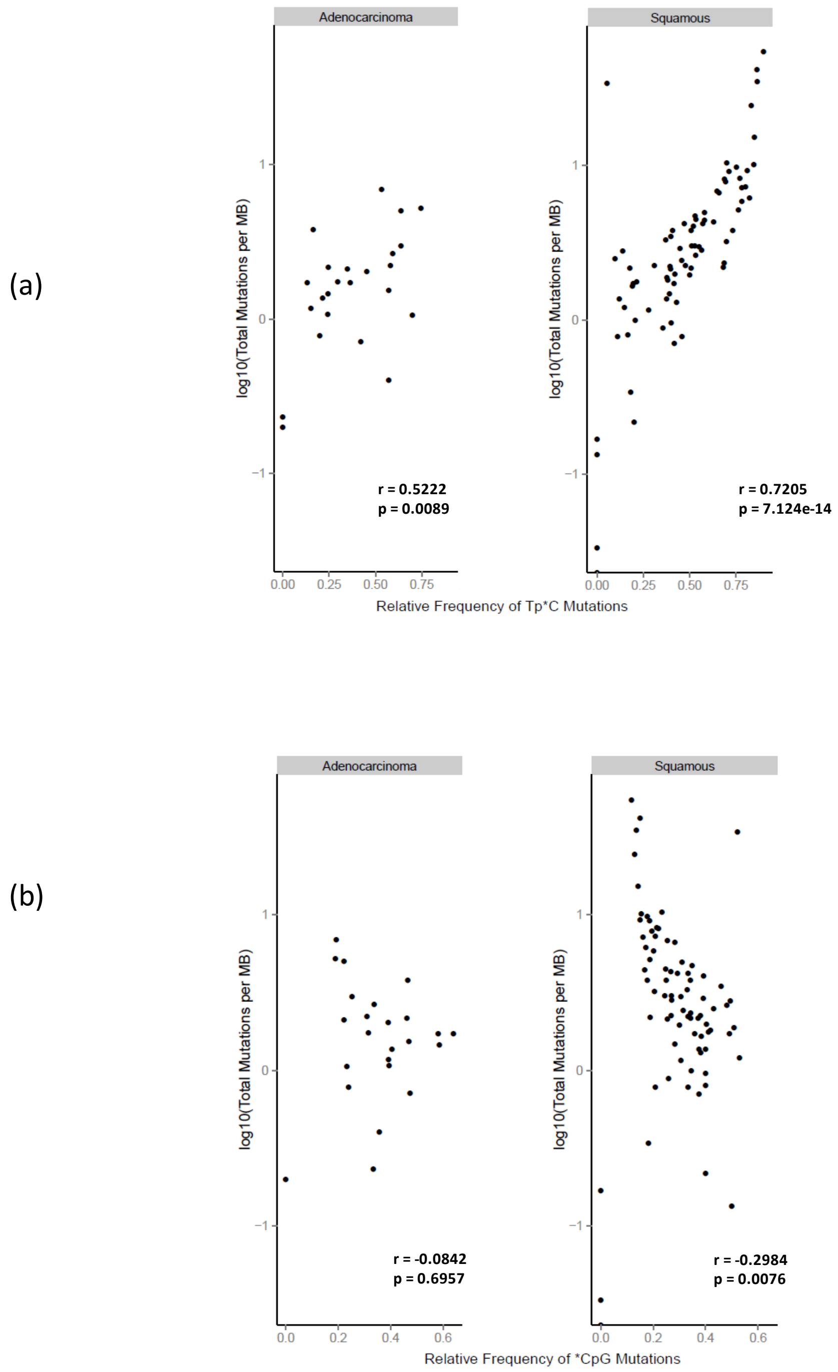


**Supplementary Figure 3. Lego plots showing mutation categories and nucleotide context** The spectrum of somatic single nucleotide variant (SNVs) is shown with the 3-base context for (a) 79 squamous cell carcinomas and (b) 24 adenocarcinomas. Colors indicate type of mutation change from reference to alternate allele where strand symmetry is folded such that all mutations are either reference C or A alleles (sequence for G and T reference alleles is reverse complemented). Each colored square shows the sequence context as labeled in the legend on the lower right. For example the tallest yellow bin in (a) represents C>T SNVs preceded by a T base and followed by an G base. The vertical scale is the rate of SNVs across all samples. The pie chart on the upper left shows the relative proportions of each SNV type. Arrows indicate the bins corresponding to the major mutational signatures for squamous (Tp*C to T/G) and adenocarcinoma (*CpG to T).



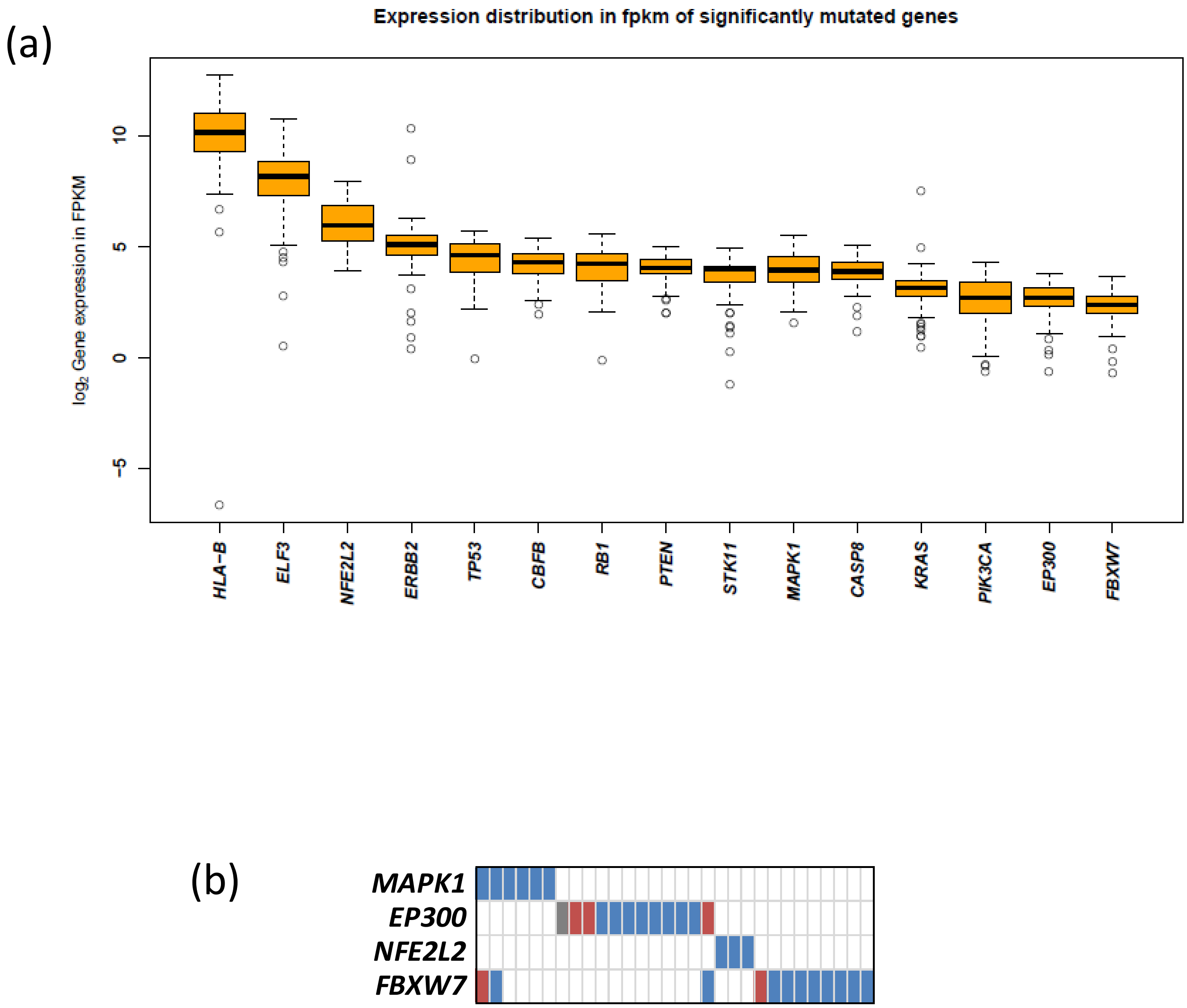
**Supplementary Figure 4. Hierarchical clustering of tumors by mutation types**

Hierarchical clustering of 115 samples by nucleotide mutational context was performed using the heatmap.2 function from the gplots library implemented in R 2.15.1. Mutation counts were scaled per sample (i.e. converted to fraction of mutations corresponding to each category) and clustered using Ward's minimum variance method. There were 5 major clusters. Clusters I and II are similar and are characterized primarily by *TpC to T/G mutations. Cluster III is characterized by tumors with few mutations spread across several mutational contexts. Cluster IV has a high relative frequency of *CpG to T mutations. Cluster V tumors have both (Tp*C and *CpG) mutational signatures. See Supplementary Note 8A for further analyses and discussion.



**Supplementary Figure 5. Relationship between mutational signatures and mutational rates**

These scatter plots show the Pearson correlation relationships between the mutational rates in tumors and the relative frequency of the (a) Tp*C and (b) *CpG mutational signatures. Each plot shows these relationships for adenocarcinomas and squamous cell carcinomas, respectively.



**Supplementary Figure 6. RNASeq-derived gene expression values and non-overlapping patterns of mutations in significantly mutated genes**

(a) The distribution of gene expression values ($\log_2$FPKM) of significantly mutated genes (q<0.1) across the whole dataset are shown in boxplots. Gene expression values were derived from TopHat-aligned RNAseq data normalized to fragments per kilobase of exon per million fragments mapped (FPKM) using Cufflinks V 2.0.2. Genes with FPKM<1 were filtered out from the list of significantly mutated genes**.**

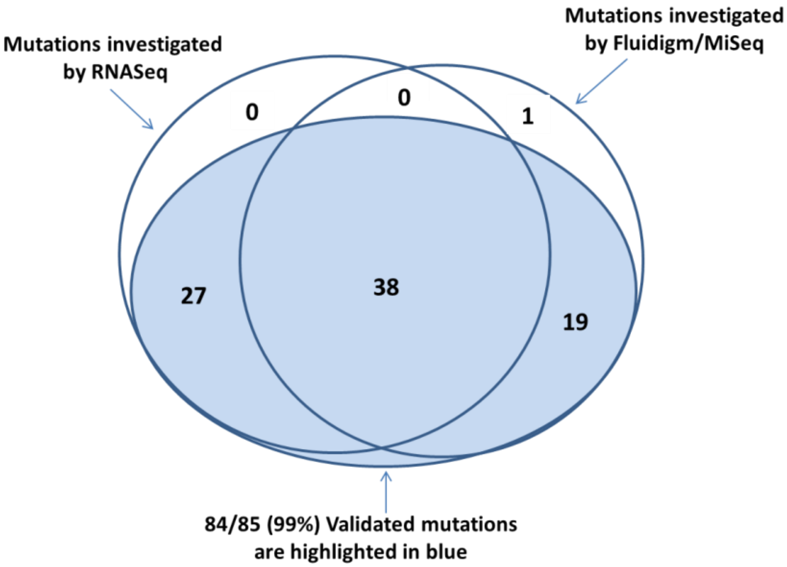
(b) The mutations in 4 of the genes reported in Figure 1 (*MAPK1, EP300, NFE2L2* and *FBXW7)* are sorted highlight the general (but not absolute) non-overlapping pattern mutations in. these 4 genes. Each vertical column represents a patient. Only the patients with at least one mutation in one of the 4 genes, are depicted. Blue bars: missense mutations, red bars: nonsense mutations, grey bars: silent mutations

## Validation of all mutations with overlapping exome, genome and transcriptome data

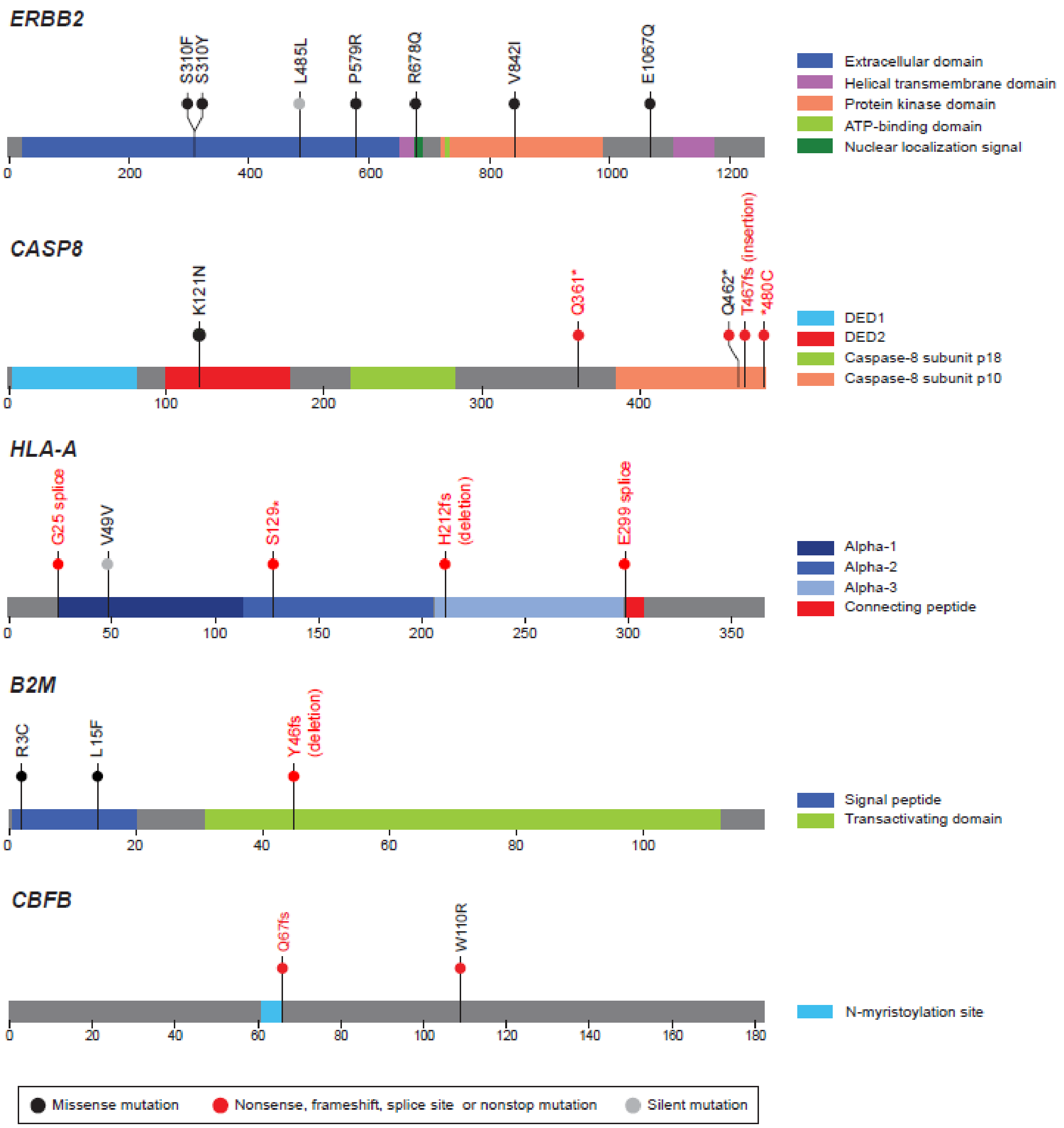
**Mutations identified by Whole Exome Sequencing**

| | Total | Powered | Validated | Validation Rate |
|---|---|---|---|---|
| RNASeq | 13346 | 5220 | 4439 | 0.85 |
| WGS | 1529 | 798 | 736 | 0.92 |
| RNASeq and WGS | 1452 | 416 | 345/415 | 0.83/.998 |

## Validation of significantly mutated genes

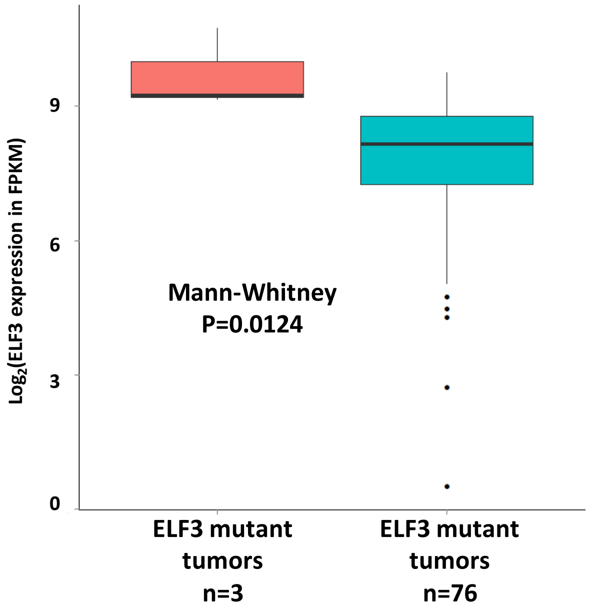

Mutations investigated by RNASeq
Mutations investigated by Fluidigm/MiSeq
0
0
1
27
38
19
84/85 (99%) Validated mutations are highlighted in blue

**Supplementary Figure 7. Validation of significantly mutated genes**
This Venn diagram shows the results of validation experiments for mutated genes. Two approaches were employed: Illumina MiSeq-based targeted resequencing of Fluidigm-derived amplicons and investigation of RNASeq data. The numbers shown represent mutations investigated by at least one method and for which there was enough power/coverage to call mutations. The counts and rate of the exome SNP mutations validated against the corresponding RNASeq and WGS files (79 and 14, respectively) are reported below the Venn diagram. The validation procedure is described in Supplementary Note 6D. For mutations investigated by both RNASeq and WGS, validation and validation rate are reported as (validated in both)/(validated in at least one) and powered is reported for powered in both.



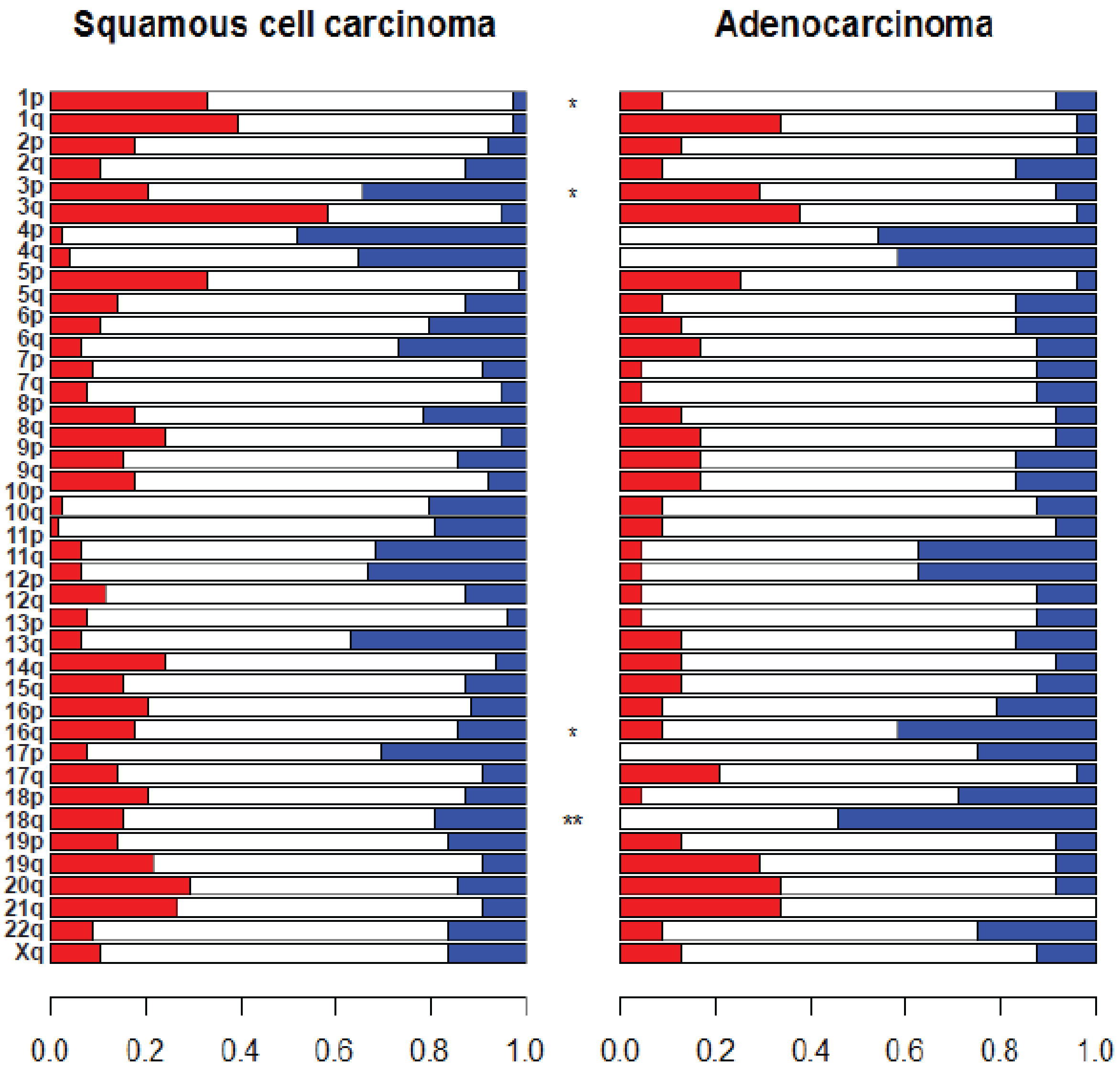
**Supplementary Figure 8. Novel somatic recurrent mutations in cervical carcinoma candidate genes**

The locations of somatic mutations in novel significantly mutated genes in cervical carcinoma: *ERBB2, CASP8, HLA-A, B2M, NFE2L2* and *CBFB* are shown in the context of protein domain models derived from UniProt and Pfam annotations. Numbers refer to amino acid residues. Each filled circle represents an individual mutated tumor sample: missense and silent mutations are represented by filled black and grey circles, respectively while nonsense, frameshift, splice site and nonstop mutations are represented by filled red circles and red text. Domains are depicted with various colors with an appropriate key located on the right hand of each domain model.



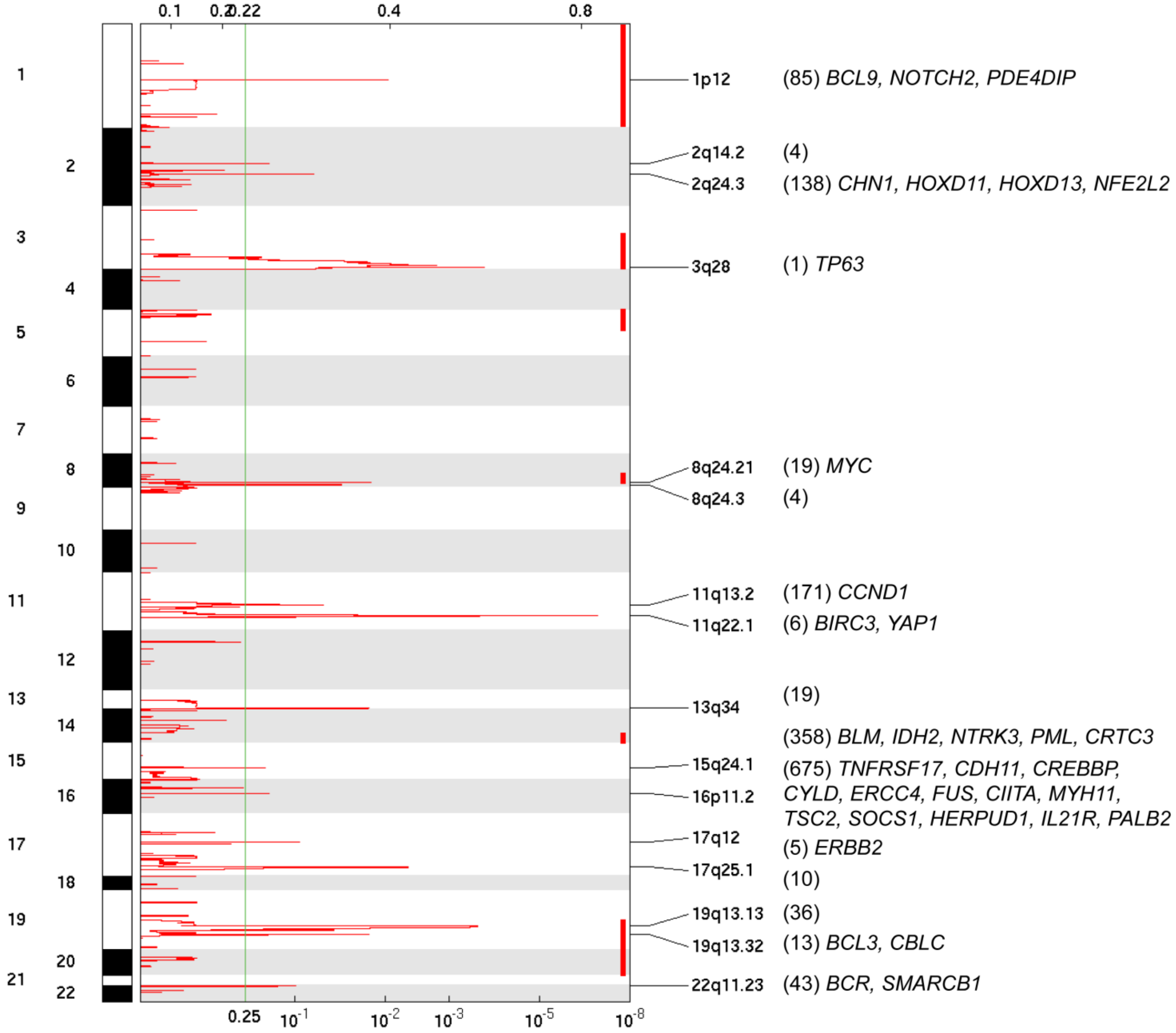
**Supplementary Figure 9. Frameshift mutations in *ELF3* are associated with high gene expression**

There were 3 tumors with *ELF3* frameshift mutations at positions 255, 330 and 350 respectively. The panel shows boxplots comparing the gene expression levels of *ELF3* in the 3 mutated tumors with gene expression levels in the 76 tumors without *ELF3* mutations.



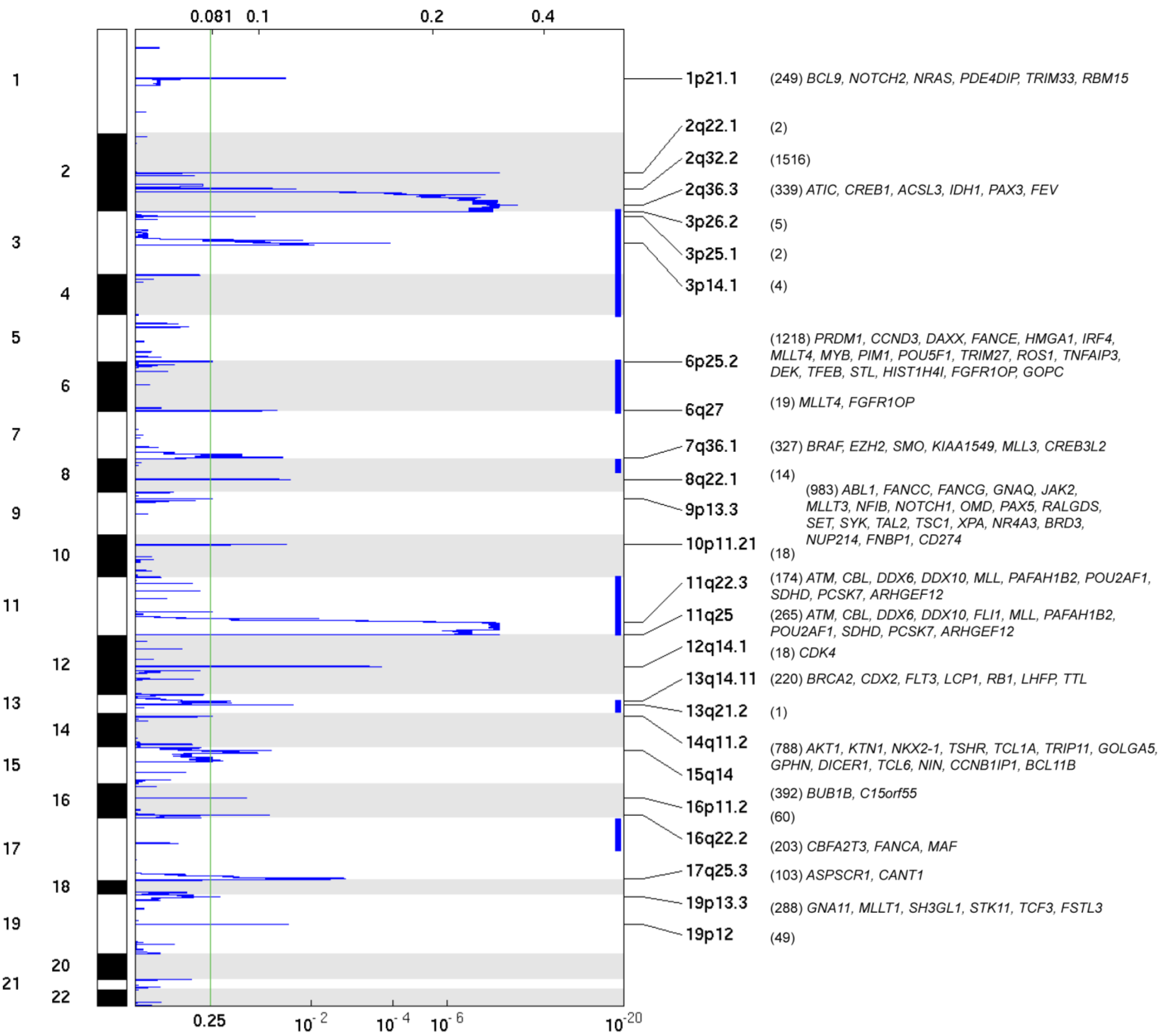
**Supplementary Figure 10. Overview of arm level copy number gains and losses in cervical squamous cell carcinomas and adenocarcinomas**

The bar graphs show the relative frequencies of arm level amplifications (red) and deletions in each chromosomal arm across 115 tumors, as determined by GISTIC 2.0 (Mermel et al, 2010). To be called an arm level alteration, amplification or deletions must encompass at least 50% of a chromosomal arm. Significant differences between the two histological types were determined by chi-squared tests.

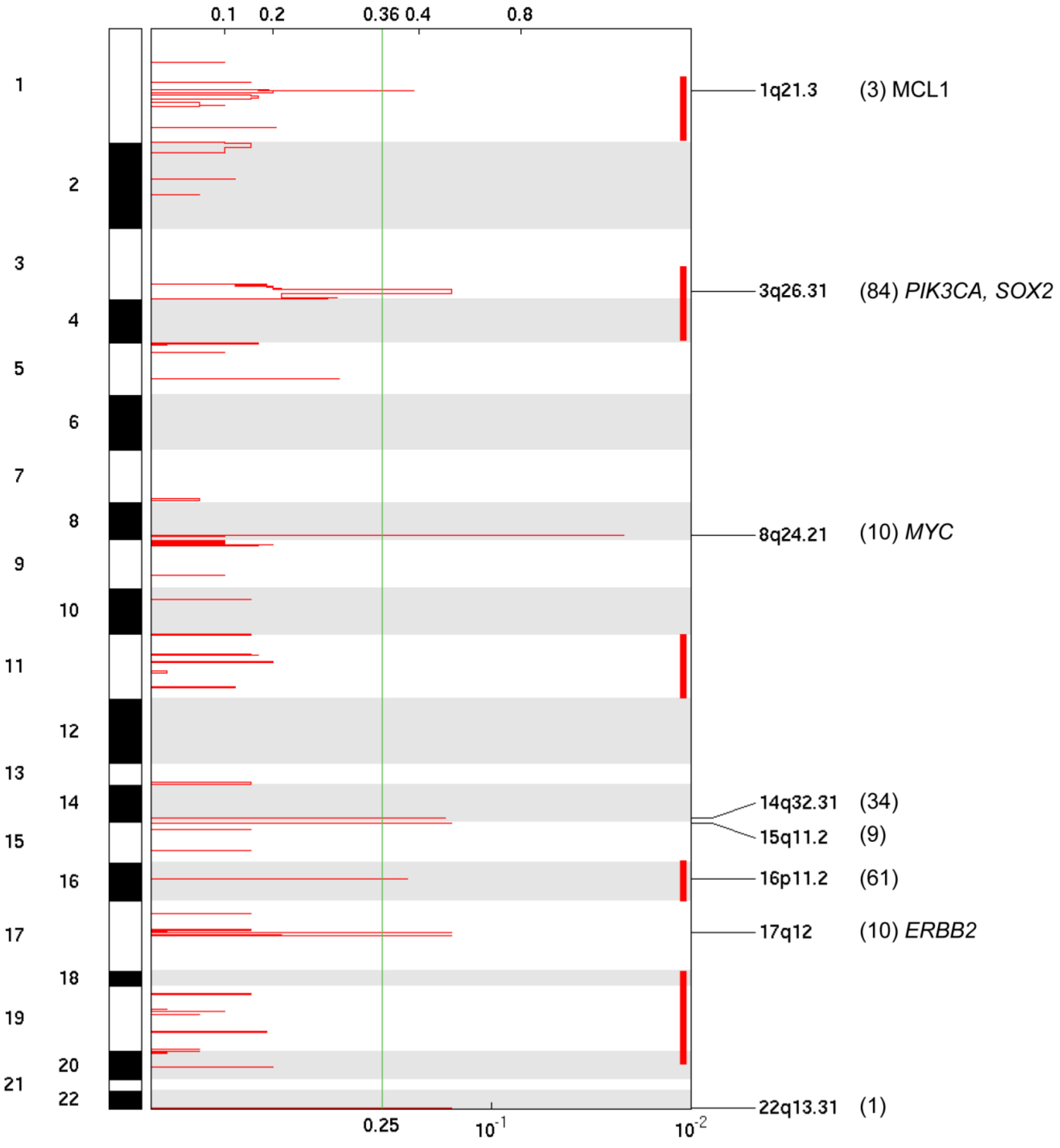


**Supplementary Figure 11. Focal amplifications across 79 cervical squamous cell carcinomas**

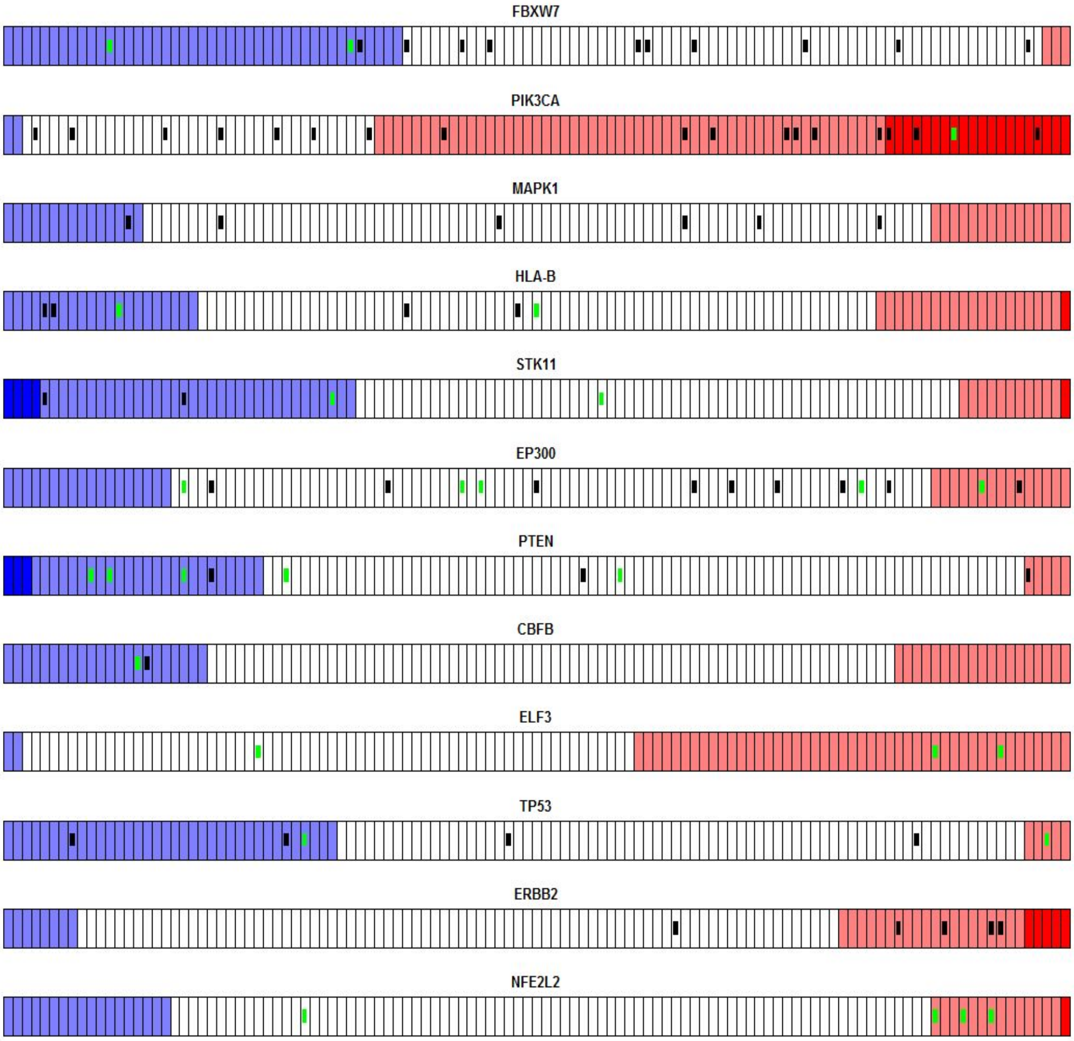
Somatic copy number alterations were analyzed by GISTIC2.0. Chromosome positions are indicated along the *y* axis. Focal amplifications are depicted with horizontal red bars, with the green line representing the significance threshold of q<0.25 (the false discovery rate after multiple hypothesis testing). The locations of the peak regions and the known cancer-related genes within those peaks are indicated to the right of each panel. The number of genes in a peak is in parentheses and the listed genes have been documented in the Cancer Gene Census. The vertical red bars indicate chromosomal arms with broad copy number gains. The vertical red bars indicate chromosomal arms with broad copy number gains.



**Supplementary Figure 12. Focal deletions across 79 cervical squamous cell carcinomas**
Somatic copy number alterations were analyzed by GISTIC2.0. Chromosome positions are indicated along the *y* axis. Focal deletions are depicted with horizontal blue bars, with the green line representing the significance threshold of q<0.25 (the false discovery rate after multiple hypothesis testing). The locations of the peak regions and the known cancer-related genes within those peaks are indicated to the right of each panel. The number of genes in a peak is in parentheses and the listed genes have been documented in the Cancer Gene Census. The vertical blue bars indicate chromosomal arms with broad copy number gains.

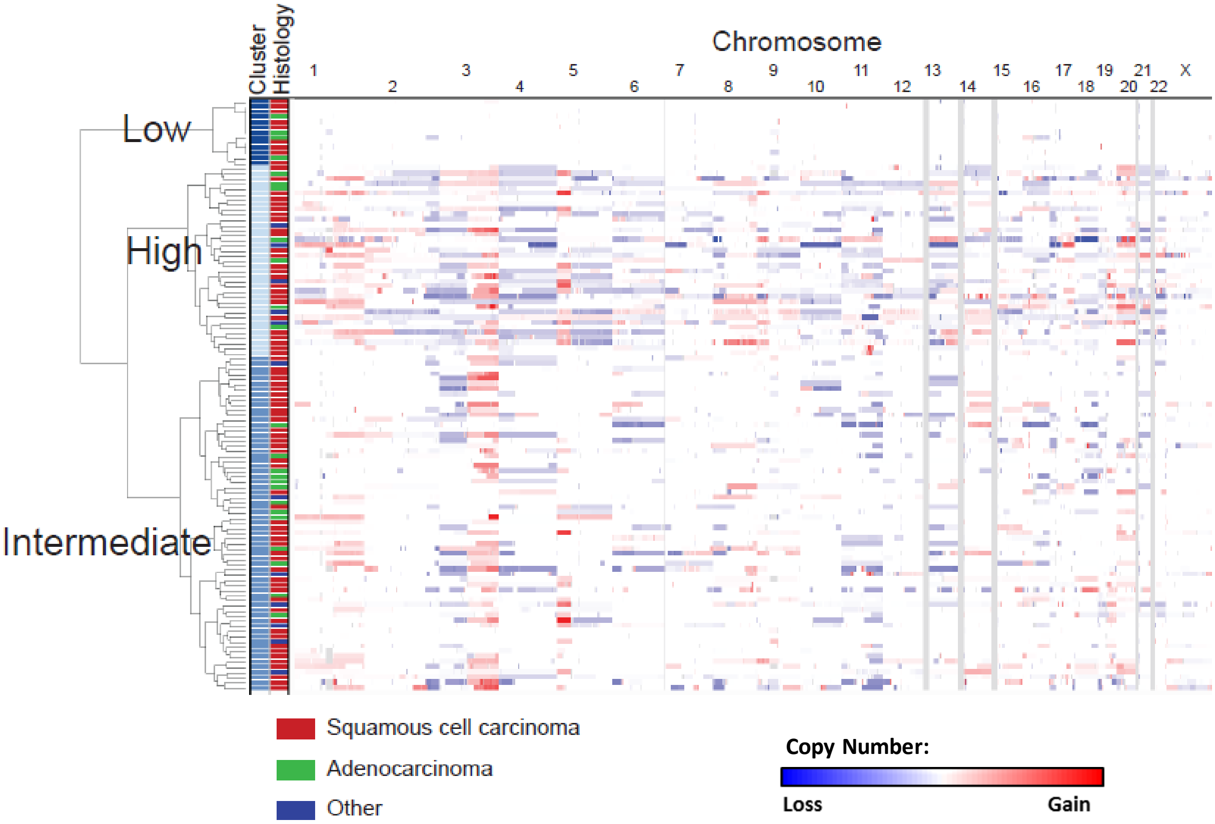


**Supplementary Figure 13. Focal amplifications across 24 cervical adenocarcinomas**
Somatic copy number alterations were analyzed by GISTIC2.0. Chromosome positions are indicated along the *y* axis. Focal amplifications are depicted with red bars, with the green line representing the significance threshold of q<0.25 (the false discovery rate after multiple hypothesis testing). The locations of the peak regions and the known cancer-related genes within those peaks are indicated to the right of each panel. The number of genes in a peak is in parentheses and the listed genes have been documented in the Cancer Gene Census. The vertical red bars indicate chromosomal arms with broad copy number gains.

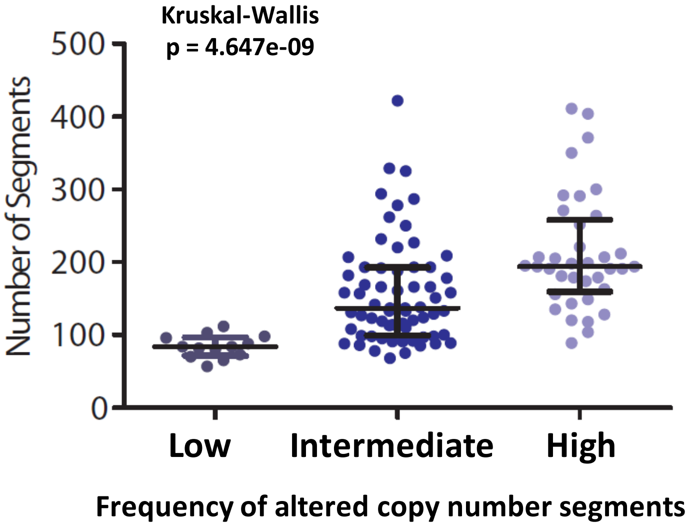


**Supplementary Figure 14. Relationship between Copy Number Alterations and Somatic Mutations**

This figure shows somatic mutation and copy number data for 115 cervical carcinoma patients, depicted in 115 contiguous bars per patient. For each gene, copy number data are sorted from the greatest degree of copy number loss on the left (blue), to the highest level of copy number gain (red) on the right. Low level copy number changes between 0.1 and 1.0 copies are represented by the lighter color shades, while the darker shades represent changes >1 copy. The white boxes represent copy number change <0.1. Somatic mutations found in each gene are superimposed upon the copy number data. Missense mutation are represented by black boxes while nonsense, frameshift and splice site mutations are represented by green boxes.

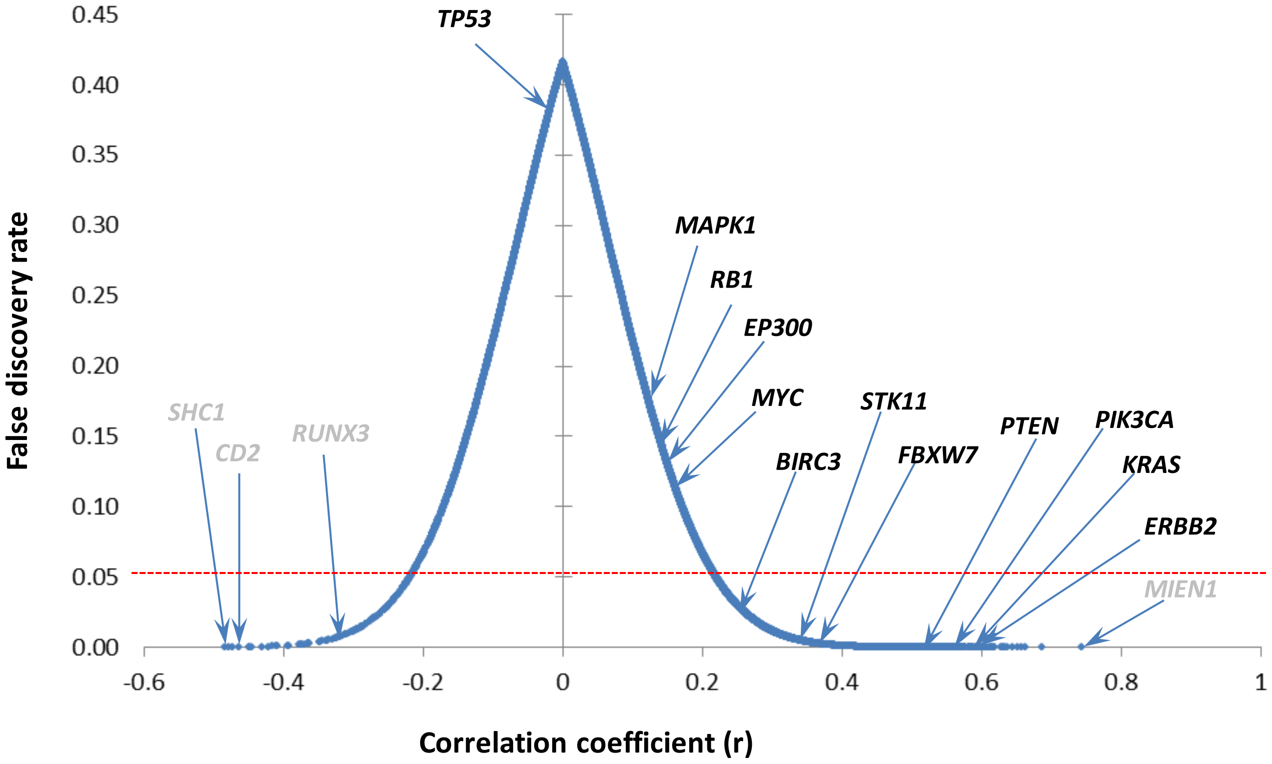


**Supplementary Figure 15. Hierarchical Clustering of Copy Number Alterations**
Hierarchical clustering of copy number data was performed on thresholded relative copy number data in significantly recurring amplification or deletion regions identified by GISTIC2.0 analysis. Copy number gains and losses are depicted in red and blue respectively. The tumors are annotated with histological data: with red, green and blue boxes representing squamous cell carcinomas, adenocarcinomas and tumors with other histological diagnoses, respectively. Three clusters were found, roughly corresponding to high, intermediate and low frequencies of copy number alterations. See Supplementary Figure 16 for statistical analyses.



**Supplementary Figure 16. Relative Frequencies of Altered Copy Number Segments in Cervical Carcinoma Copy Number Hierarchical Clusters**

After hierarchical clustering was performed on the copy number profiles of the tumors, three groups were identified. These groups (low, intermediate and high) differed significantly in their number of copy number segments. The number of copy number segments for tumors in each group are plotted above.



**Supplementary Figure 17. Correlation between copy number alterations and gene expression in cervical carcinomas**

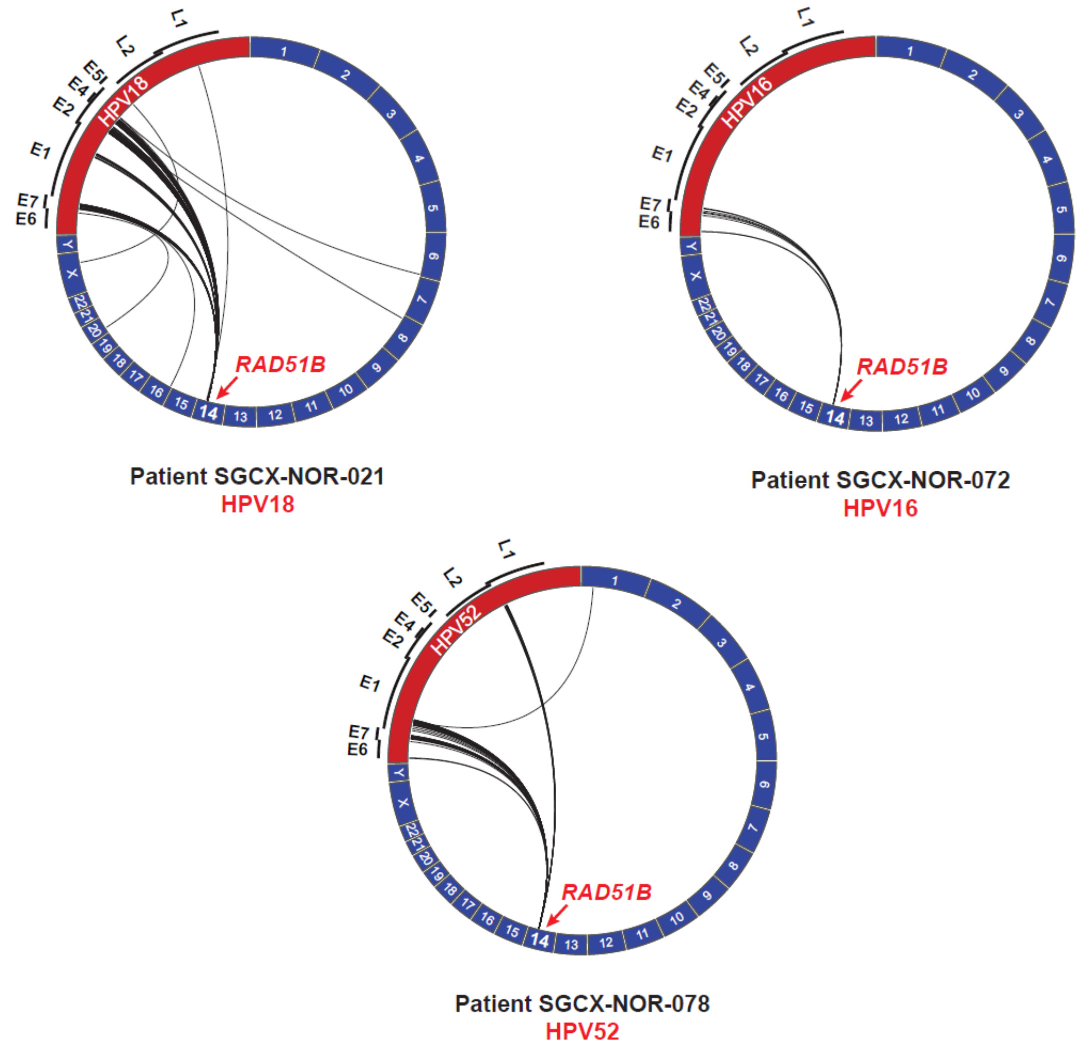
We performed "genome-wide" Pearson correlation analysis between DNA copy number changes and RNASeq-derived gene expression (FPKM) for all tumors with both sets of data available, and calculated false discovery rates (q) for each correlation. This figure shows correlation coefficients on the X axis and the corresponding q values on the Y axis, for 16898 genes. Only 6874 (41%) of the genes investigated had significant correlative relationships at a false discovery rate of q<0.05, with the lowest corresponding correlation coefficient (r) values being +0.22 and -0.22. Some of the significantly mutated or altered genes in this study such as *ERBB2, KRAS, PIK3CA, PTEN, FBXW7, STK11* and *BIRC3* had r values greater than 0.22. Other genes including *TP53, MYC, EP300, RB1* and *MAPK1* had r<0.22, with q values above 0.1. *TP53* showed no correlation between copy number and gene expression with r=-0.03 and q=0.36.

| f-HPV | MassArray | | Positive in RNASeq | Negative in RNASeq | RNA not available |
|---|---|---|---|---|---|
| Positive | Positive | 99 | 69 | 4 | 26 |
| Positive | Negative | 8 | 4 | 1 | 3 |
| Negative | Positive | 2 | | | 2 |
| Negative | Negative | 2 | | | 2 |
| Positive | N/A | 2 | | | 2 |
| N/A | Positive | 1 | 1 | | |
| N/A | N/A | 1 | | | 1 |
| | | 115 | 74 | 5 | 36 |

*N/A: Sample not adequate/available for test*

**Supplementary Figure 18. HPV typing data by method**

This shows a comparison of 3 methods used to assess HPV status: 2 multiplex HPV DNA PCR-based methods, f-HPV and MassArray, as well as RNA sequencing. One hundred and nine (96%) of 113 tumors were found to be positive by the f-HPV test, while 102 (91%) of 112 tumors were found to be HPV-positive by the MassArray method. Of the 111 tumors with data available from both tests, there was agreement in 101 (91%) tumors for HPV positivity. Seventy-three (94%) of 78 tumors with both f-HPV and RNASeq data were HPV-positive by both methods. Similarly, 69 (95%) of 73 tumors with both MassArray and RNASeq data were HPV positive by both methods.

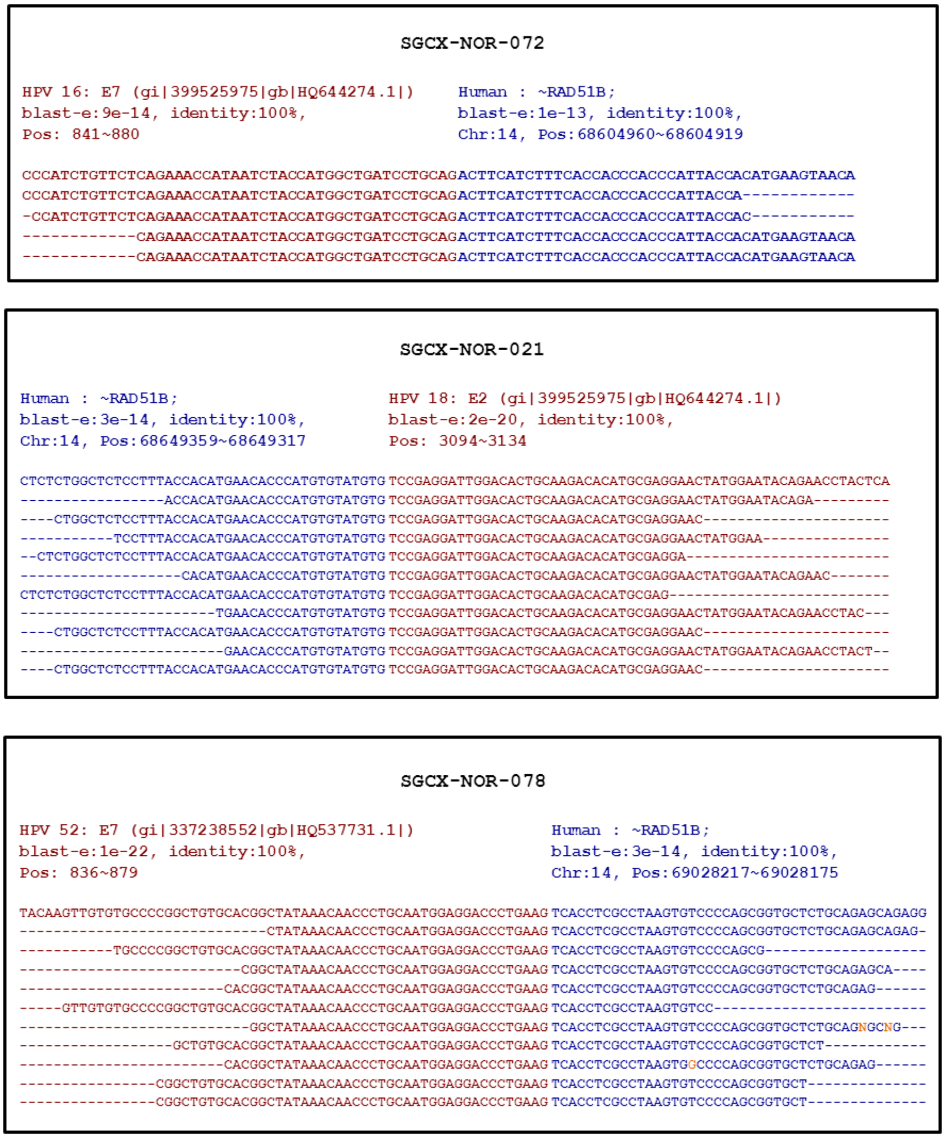


Patient SGCX-NOR-021
HPV18

Patient SGCX-NOR-072
HPV16

Patient SGCX-NOR-078
HPV52

**Supplementary Figure 19. Recurrent HPV integration into the *RAD51B* locus**
The Circos plots depict *RAD51B* integration events present in 3 of the 79 tumors with transcriptome sequencing data. In each case, there was a different viral strain (HPV16, HPV18 and HPV52 in samples SGCX-NOR-072, SGCX-NOR-021, SGCX-NOR-078 respectively) integrated within the *RAD51B* locus on chromosome 14. The lines within these Circos plots display the locations of chimeric read pairs in which one pair mate is human and the other pair mate is viral. For example, in sample SGCX-NOR 021, the reads originate from spliced E6, E7, E1 transcripts and the integration breakpoint is near the end of E2. In addition, there is a read on the other side of the integration event represented by the chimeric read in L1.

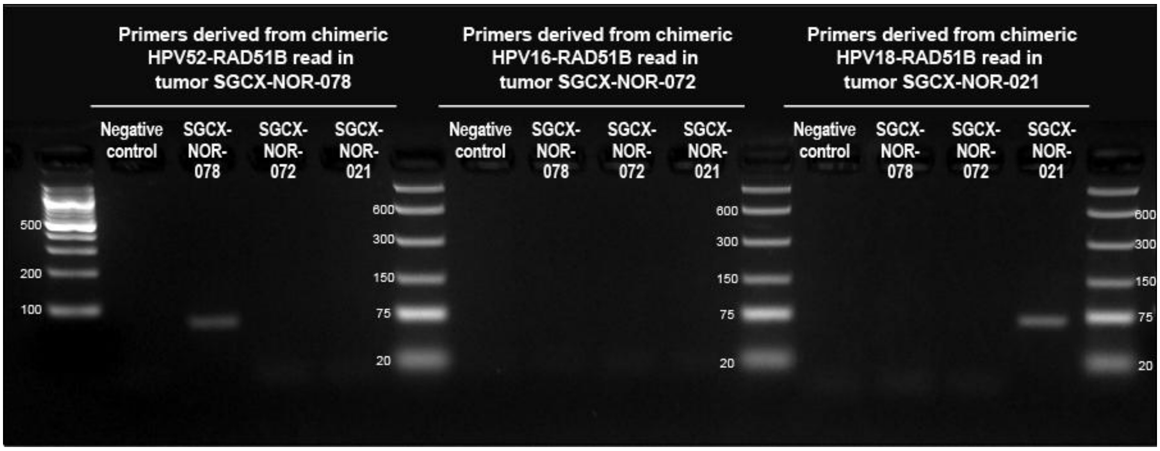


```
SGCX-NOR-072

HPV 16: E7 (gi|399525975|gb|HQ644274.1|)     Human : ~RAD51B;
blast-e:9e-14, identity:100%,                blast-e:1e-13, identity:100%,
Pos: 841~880                                 Chr:14, Pos:68604960~68604919

CCCATCTGTTCTCAGAAACCATAATCTACCATGGCTGATCCTGCAGACTTCATCTTTCACCACCCACCCATTACCACATGAAGTAACA
CCCATCTGTTCTCAGAAACCATAATCTACCATGGCTGATCCTGCAGACTTCATCTTTCACCACCCACCCATTACCA------------
-CCATCTGTTCTCAGAAACCATAATCTACCATGGCTGATCCTGCAGACTTCATCTTTCACCACCCACCCATTACCAC-----------
------------CAGAAACCATAATCTACCATGGCTGATCCTGCAGACTTCATCTTTCACCACCCACCCATTACCACATGAAGTAACA
------------CAGAAACCATAATCTACCATGGCTGATCCTGCAGACTTCATCTTTCACCACCCACCCATTACCACATGAAGTAACA
```

```
SGCX-NOR-021

Human : ~RAD51B;                          HPV 18: E2 (gi|399525975|gb|HQ644274.1|)
blast-e:3e-14, identity:100%,             blast-e:2e-20, identity:100%,
Chr:14, Pos:68649359~68649317             Pos: 3094~3134

CTCTCTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAATACAGAACCTACTCA
-----------------ACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAATACAGA---------
----CTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAAC----------------------
-----------TCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAA---------------
--CTCTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGA------------------------
-------------------CACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAATACAGAAC-------
CTCTCTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAG--------------------------
-----------------------TGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAATACAGAACCTAC---
----CTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAAC----------------------
------------------------GAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAACTATGGAATACAGAACCTACT--
----CTGGCTCTCCTTTACCACATGAACACCCATGTGTATGTG TCCGAGGATTGGACACTGCAAGACACATGCGAGGAAC----------------------
```

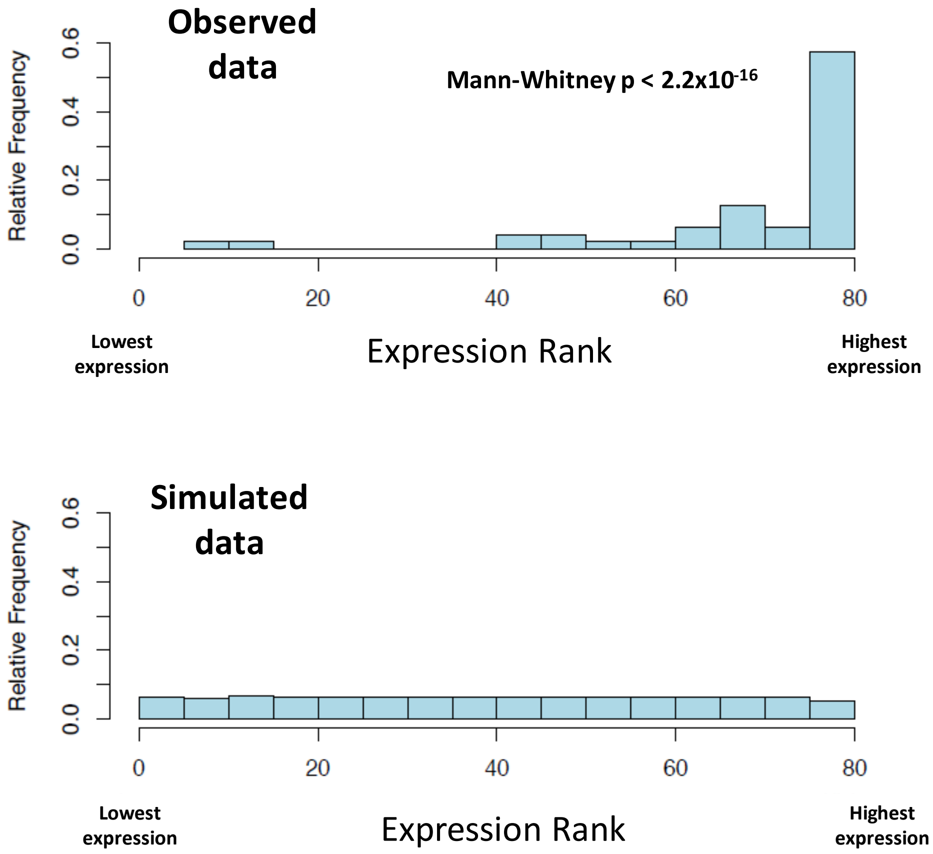
```
SGCX-NOR-078

HPV 52: E7 (gi|337238552|gb|HQ537731.1|)                    Human : ~RAD51B;
blast-e:1e-22, identity:100%,                               blast-e:3e-14, identity:100%,
Pos: 836~879                                                Chr:14, Pos:69028217~69028175

TACAAGTTGTGTGCCCCGGCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCTGCAGAGCAGAGG
-----------------------------CTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCTGCAGAGCAGAG-
-----------TGCCCCGGCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCG-------------------
--------------------------CGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCTGCAGAGCA----
------------------------CACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCTGCAGAG------
-----GTTGTGTGCCCCGGCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCC-------------------------
---------------------------GGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCTGCAGNGCNG---
------------------GCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCTCT------------
------------------------CACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGGCCCCAGCGGTGCTCTGCAGAG------
----------------CGGCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCT--------------
----------------CGGCTGTGCACGGCTATAAACAACCCTGCAATGGAGGACCCTGAAG TCACCTCGCCTAAGTGTCCCCAGCGGTGCT--------------
```

**Supplementary Figure 20. Human-Viral Fusion Reads Involving the RAD51B locus**
In addition to detecting integration sites by using split reads, we identified reads supporting RNA fusions in which single reads contained partly human and partly viral RNA. Here, fusion reads are shown for each of the RAD51B integration sites in individuals SGCX-NOR-072, SGCX-NOR-021 and SGCX-NOR-078 involving HPV16, HPV18 and HPV52 respectively. The red base pairs are viral RNA and the blue base pairs are human. Orange bases represent mismatches with the human reference genome. Fusion reads were detected through blast alignments to databases containing human and viral sequences.


Primers derived from chimeric HPV52-RAD51B read in tumor SGCX-NOR-078
Primers derived from chimeric HPV16-RAD51B read in tumor SGCX-NOR-072
Primers derived from chimeric HPV18-RAD51B read in tumor SGCX-NOR-021
Negative control
SGCX-NOR-078
SGCX-NOR-072
SGCX-NOR-021
Negative control
SGCX-NOR-078
SGCX-NOR-072
SGCX-NOR-021
Negative control
SGCX-NOR-078
SGCX-NOR-072
SGCX-NOR-021
500
200
100
600
300
150
75
20
600
300
150
75
20
600
300
150
75
20

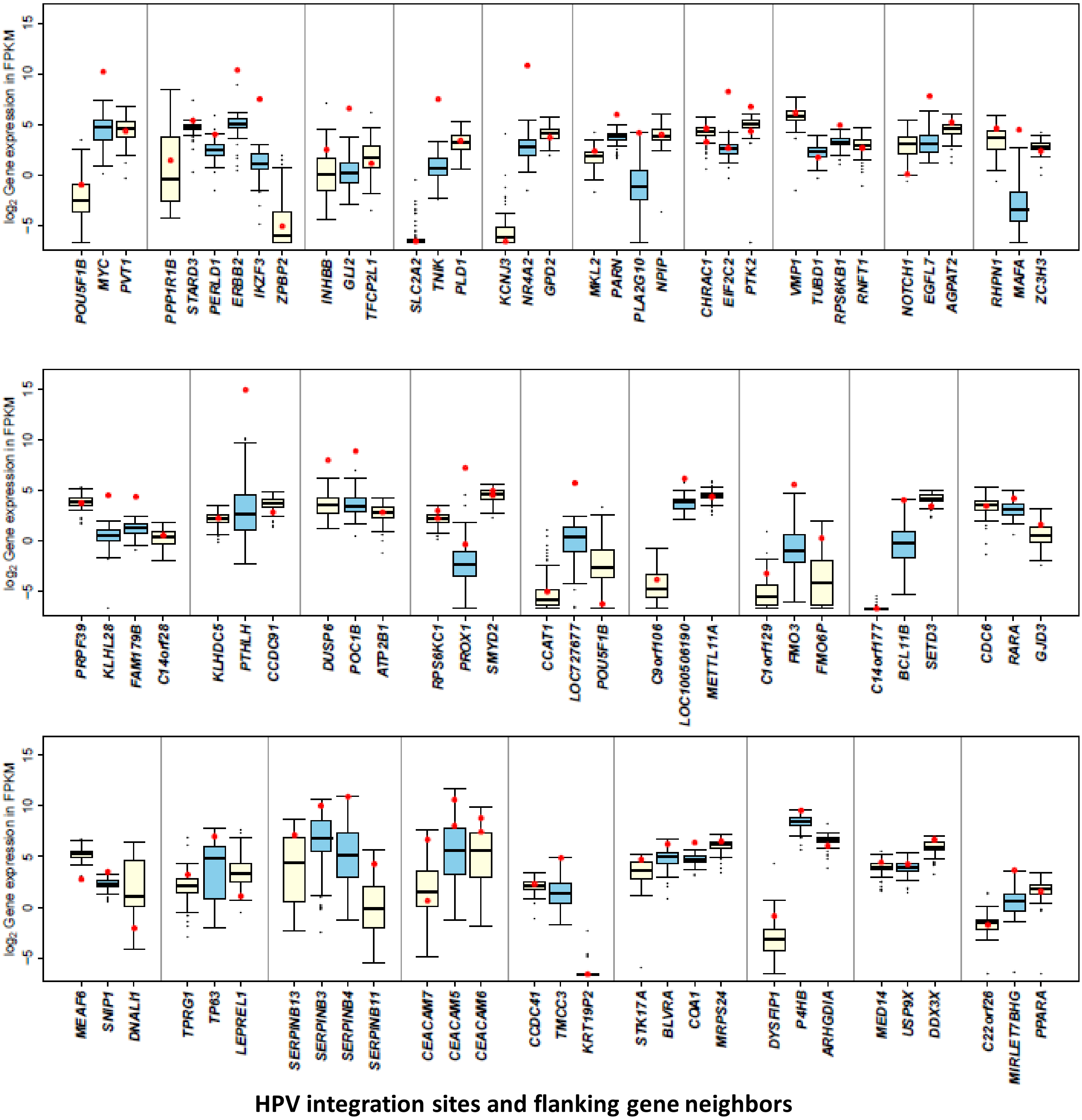
**Supplementary Figure 21. RT-PCR Validation of Human-Viral Fusion Reads Involving the RAD51B locus**

In order to validate the presence of HPV integration sites in RAD51B in three tumors, PCR primers were designed which targeted RAD51B-HPV fusion regions. RNA was isolated from the three tumors, and RT-PCR was positive for two of the three integration sites (SGCX-NOR-078 and SGCX-NOR-021). The third integration site (HPV16-RAD51B in tumor SGCX-NOR-072) was supported by the lowest number of reads in the RNA-Seq data.

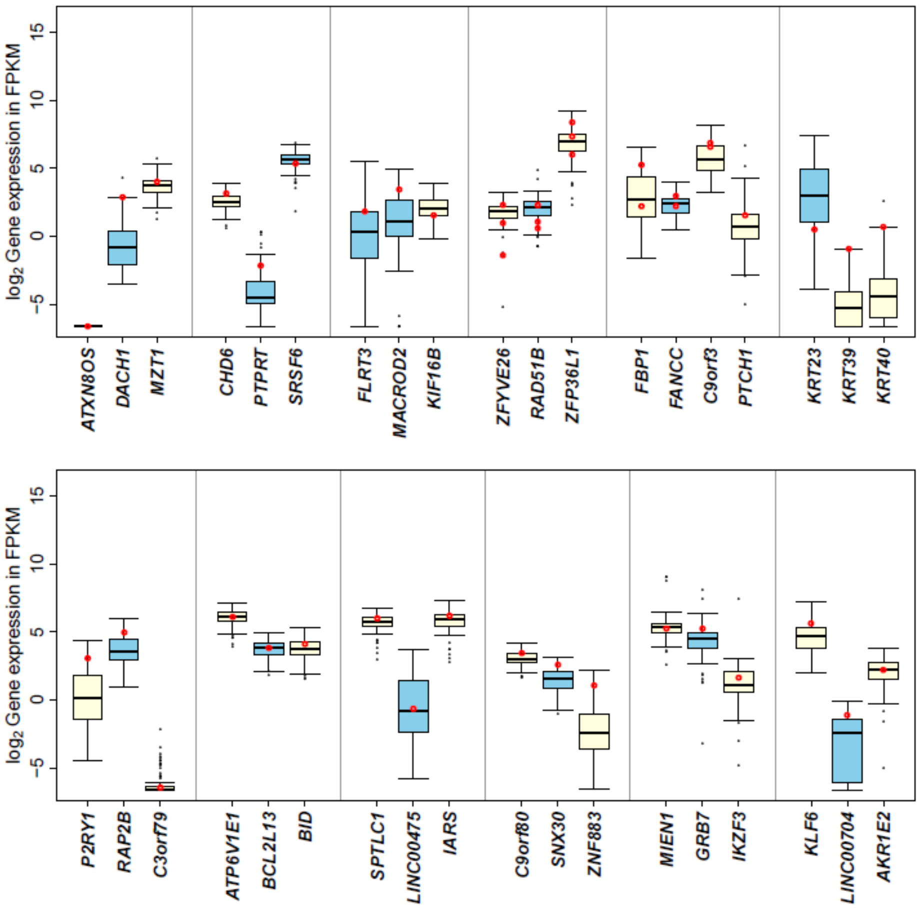


**Supplementary Figure 22. The Expression of HPV Integration Sites Genes Rank Higher in Tumors With Integration Than Tumors Without Integration**

This figure displays histograms of observed and simulated data on the ranks of gene expression levels in genes involved in HPV integration across 79 tumors. We observed that genes with HPV-human chimeric read pairs tended to have high expression in the tumor with integration as compared to all other tumors (top histogram). The bottom histogram was obtained by 10,000 separate samplings of ranks without replacement and shows the distribution of ranks expected by chance. The distribution of ranks in the simulated data is not uniform due to ties in the true ranks. See Figure 3b and Supplementary Note 11E for more details.

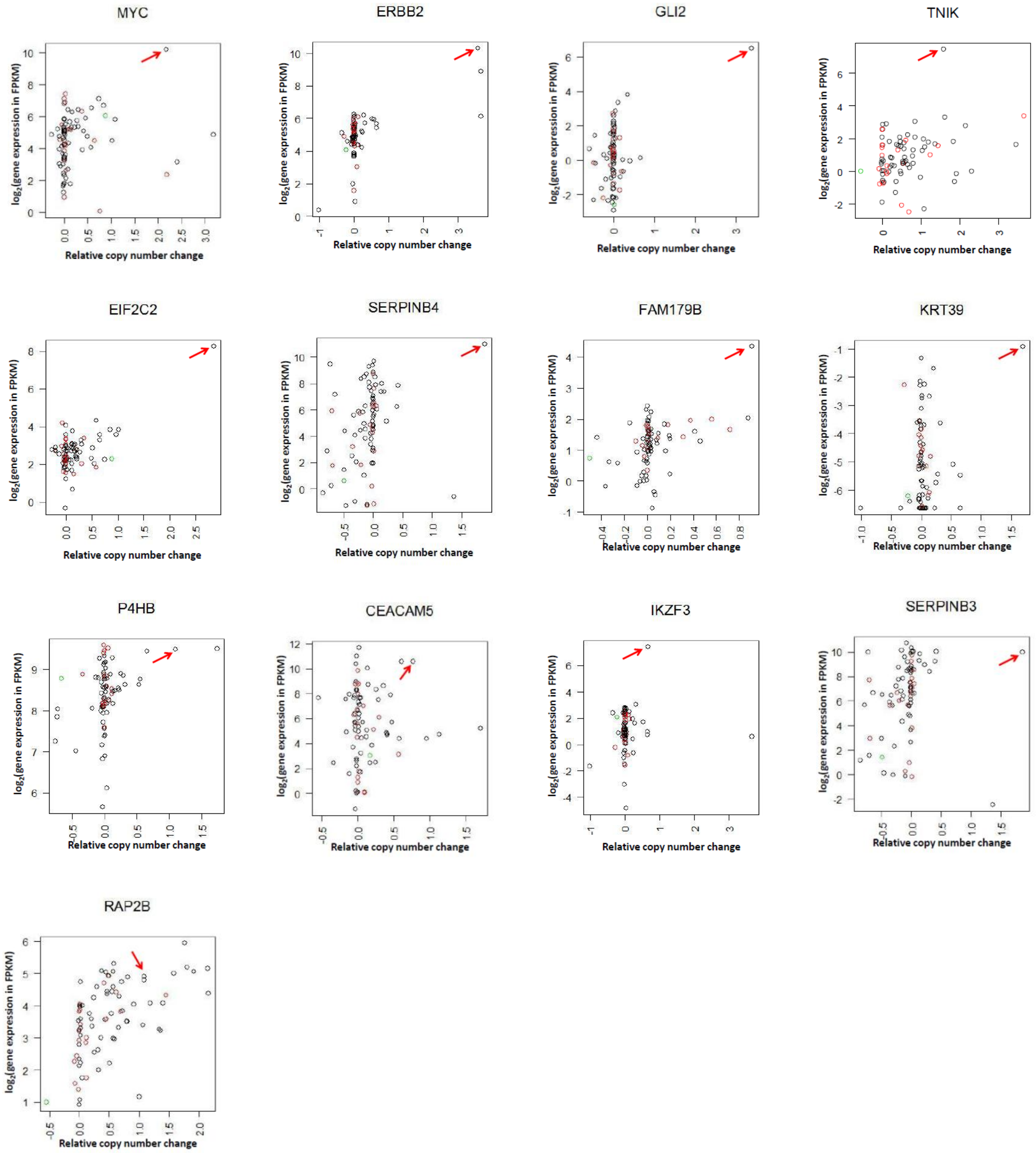


**Supplementary Figure 23. Effect of exonic HPV integration on local gene expression**
The distribution of expression levels (log$_2$FPKM) for genes near genomic HPV integration sites are shown in box plots, each across 79 tumors. In each group, the blue boxplots represent human genes with chimeric human-HPV reads pairs, which have reads mapping to exons within that gene. The flanking yellow boxplots represent the immediate 5' and 3' genomic neighbors of the integration sites. Some integration events involve multiple contiguous genes within the same locus and are represented by multiple blue boxplots. Expression levels for the integration site genes in the respective tumors with HPV integration are highlighted with red circles. Genes with recurrent integration have multiple red circles in each boxplot. See Supplementary Note 11F.



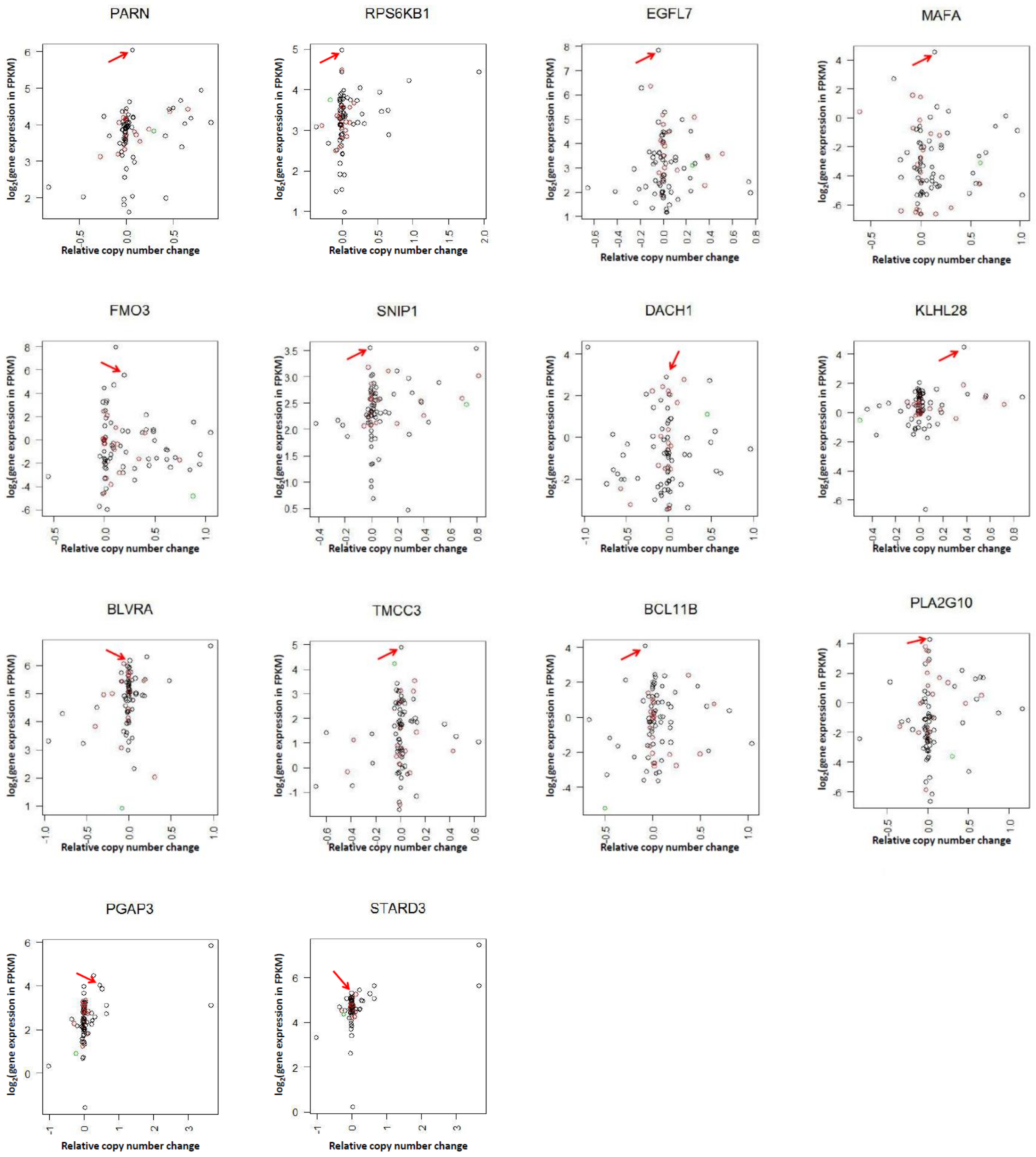
**HPV integration sites and flanking gene neighbors**

**Supplementary Figure 24. Effect of intronic HPV integration on local gene expression**
The distribution of expression levels ($\log_2$FPKM) for genes near genomic HPV integration sites are shown in box plots, each across 79 tumors. In each group, the blue boxplots represent human genes with chimeric human-HPV reads pairs, which have reads mapping to introns within that gene. The flanking yellow boxplots represent the immediate 5’ and 3’ genomic neighbors of the integration sites. Some integration events involve multiple contiguous genes within the same locus and are represented by multiple blue boxplots. Expression levels for the integration site genes in the respective tumors with HPV integration are highlighted with red circles. Genes with recurrent integration have multiple red circles in each boxplot. See Supplementary Note 11F.



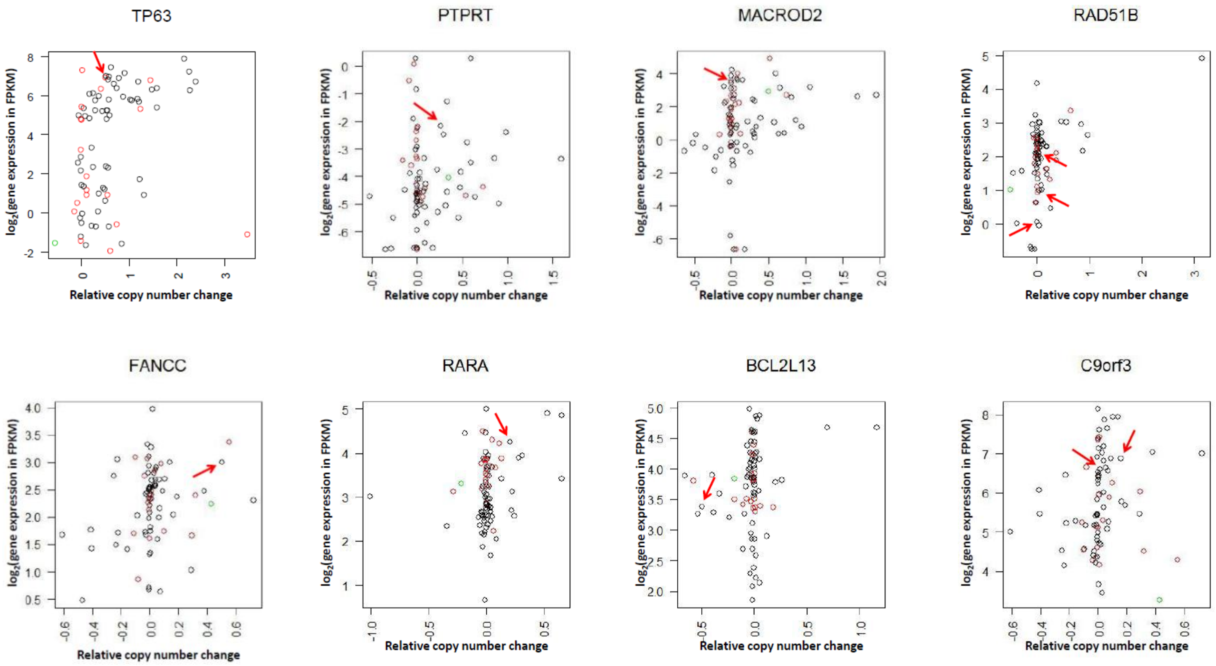
**Supplementary Figure 25. Integration site genes with high gene expression and copy number gain**

The scatter plots above compare copy number alterations and gene expression levels across 79 tumors in integration site genes with high gene expression and copy number gain. The red arrows indicates the values for tumors with  HPV integration events involving the respective genes.



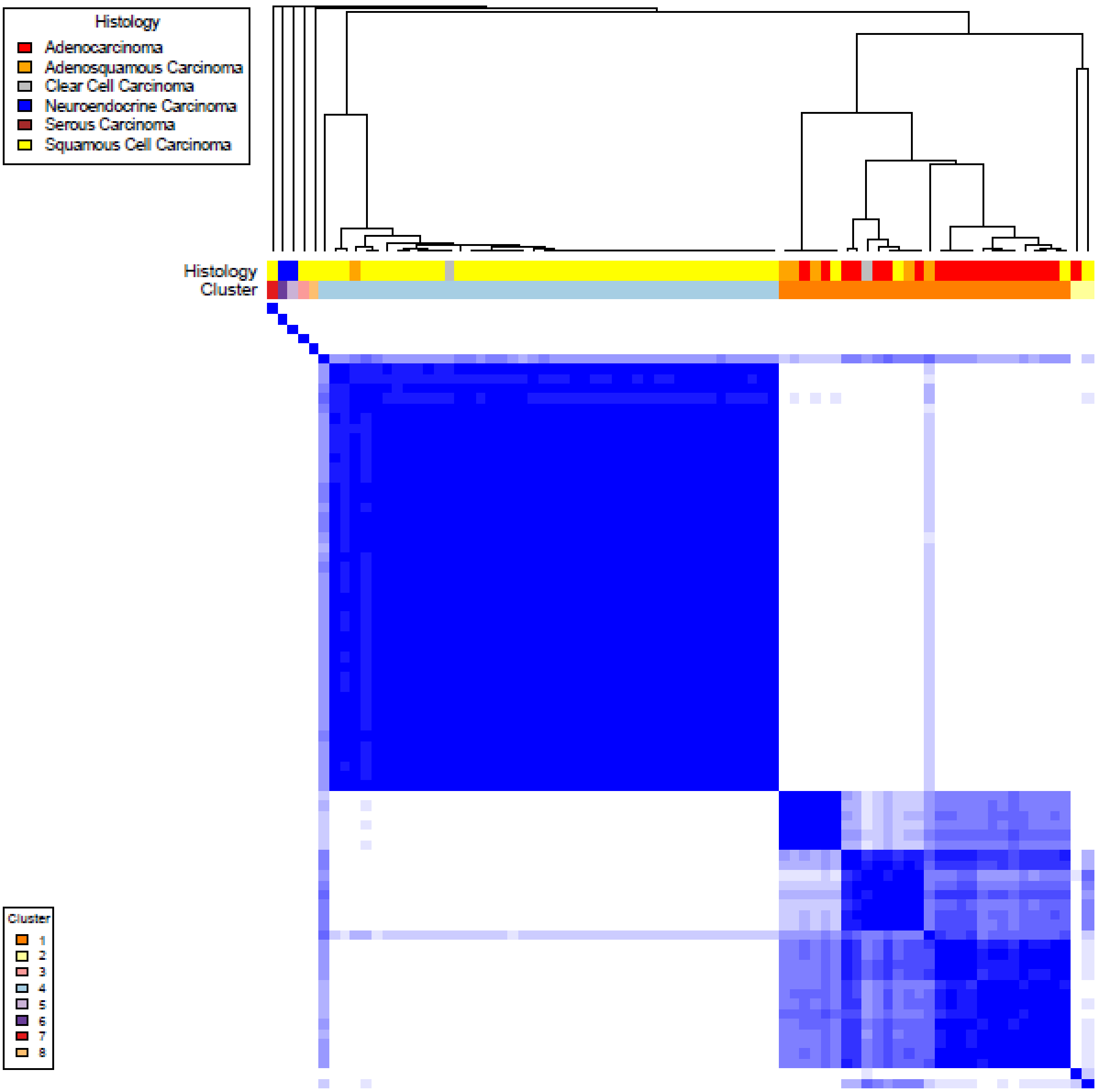
**Supplementary Figure 26. Integration site genes with high gene expression without copy number alterations**

The scatter plots above compare copy number alterations and gene expression levels across 79 tumors in integration site genes with high gene expression without copy number alterations. The red arrows indicates the values for tumors with HPV integration events involving the respective genes.



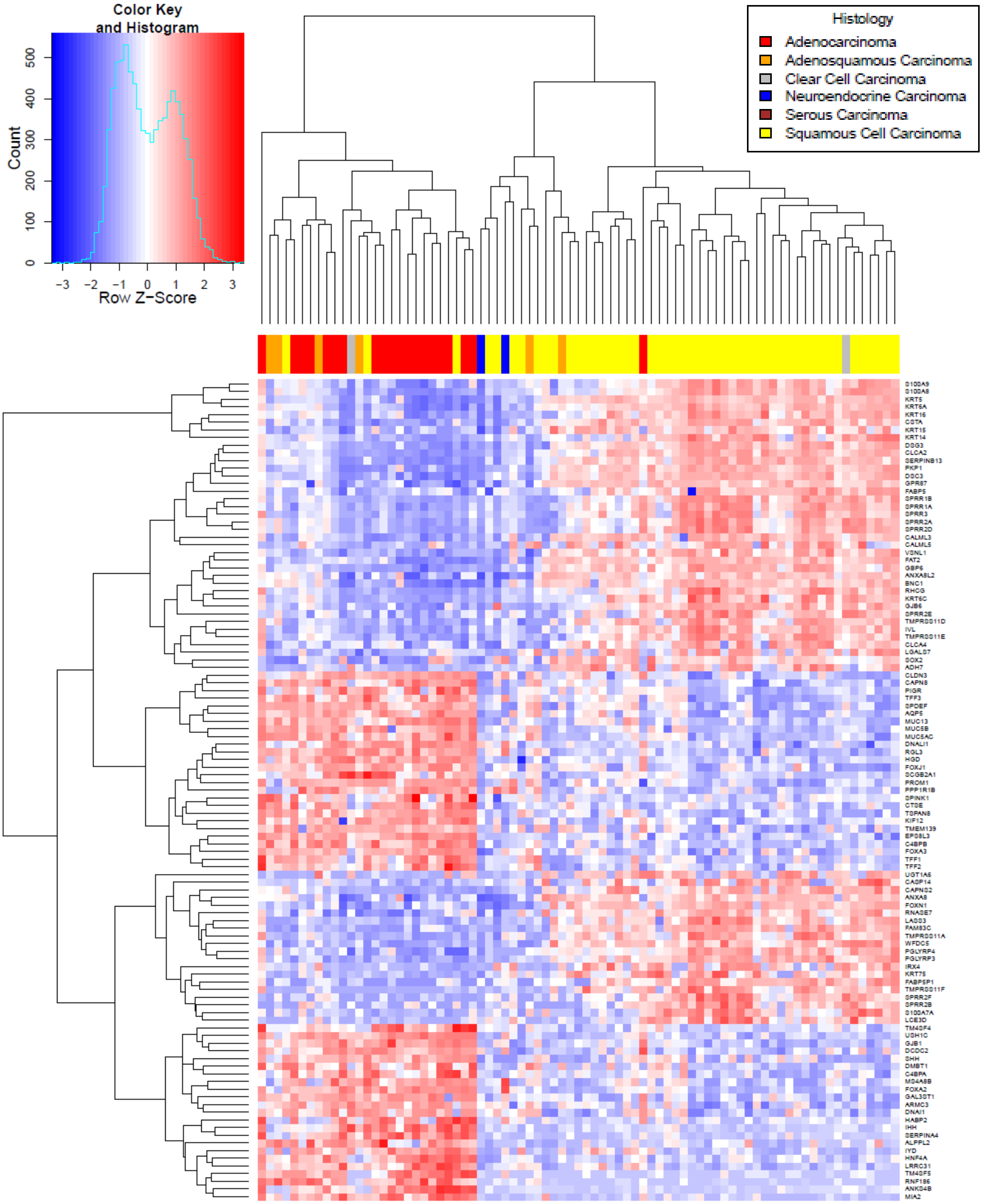
**Supplementary Figure 27. Integration site genes with miscellaneous gene expression and copy number relationships**

The scatter plots above compare copy number alterations and gene expression levels across 79 tumors in integration site genes with miscellaneous gene expression and copy number relationships. The red arrows indicates the values for tumors with HPV integration events involving the respective genes.



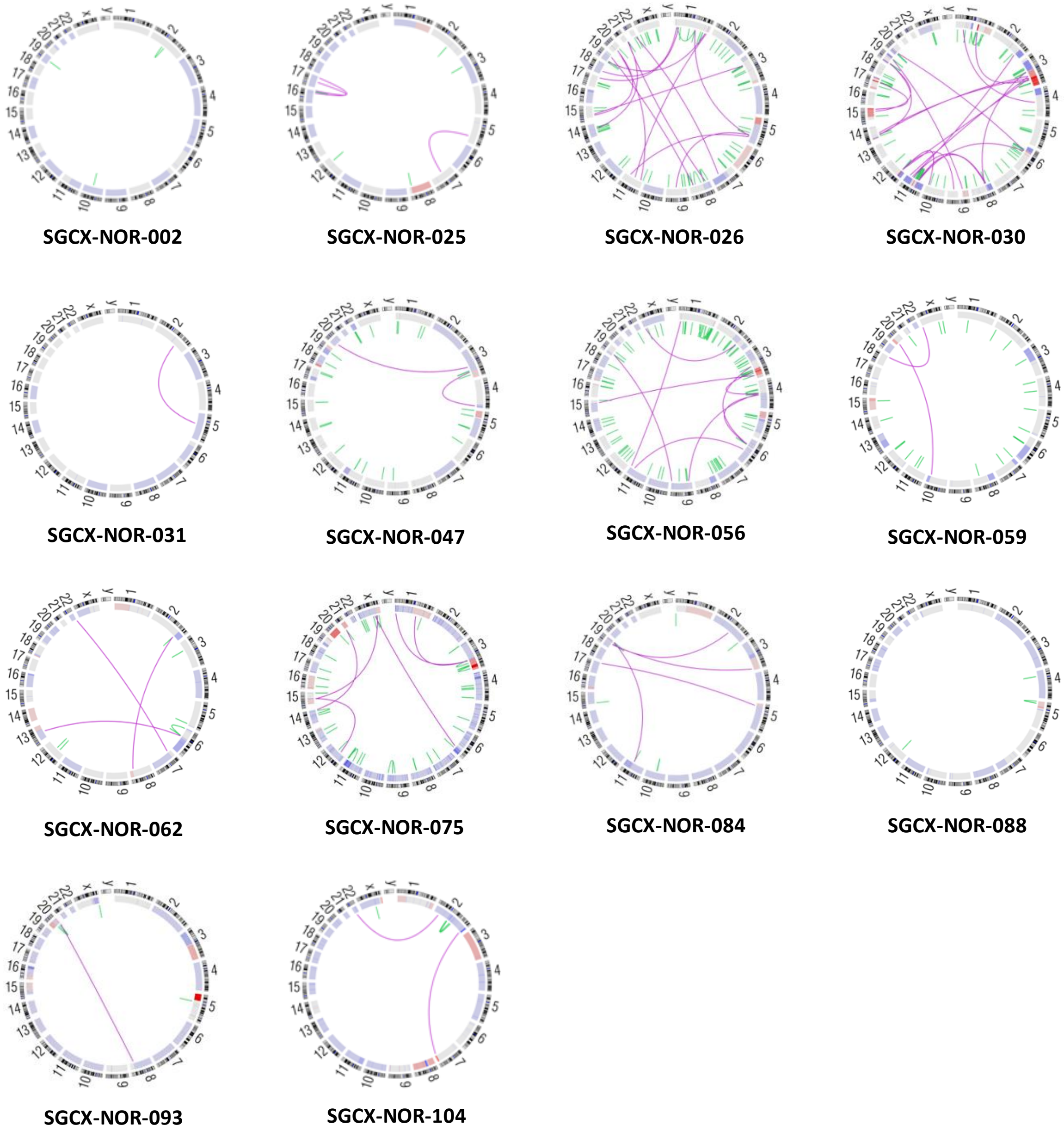
**Supplementary Figure 28. Hierarchical clustering of Gene expression data from 79 cervical carcinomas**

Gene expression ($\log_2$(FPKM)) variability was assessed in terms of the median absolute deviation across patients. The 5000 genes with the largest deviation were selected for clustering. ConsensusClusterPlus analysis (Wilkerson and Hayes. Bioinformatics 2005) was performed using 1000 resampling iterations and a maximum of 25 clusters. A k of 8 was chosen due to a clear bimodal behavior, associated with little change in the area under the empirical cumulative distribution, upon further increases in k. The color bars show histology and HPV status of each sample.



**Supplementary Figure 29. Differential gene expression in squamous cell carcinomas vs adenocarcinomas of the cervix**

This heatmap represents the differential expression between the member genes of the two largest consensus clusters in Supplementary Figure 30. The rows represent the 107 genes with > 4-fold change between the highest and lowest expression values or a q value of <0.01. Columns represent individual samples. The histological diagnosis for each tumor is shown in the top bar.



SGCX-NOR-002 SGCX-NOR-025 SGCX-NOR-026 SGCX-NOR-030

SGCX-NOR-031 SGCX-NOR-047 SGCX-NOR-056 SGCX-NOR-059

SGCX-NOR-062 SGCX-NOR-075 SGCX-NOR-084 SGCX-NOR-088

SGCX-NOR-093 SGCX-NOR-104

**Supplementary Figure 30. Spectrum of genomic rearrangements across 14 cervical carcinoma tumors**

In these Circos plots, chromosomes are arranged circularly end-to-end with each chromosome's cytobands marked in the outer ring. The inner ring displays copy number data inferred from whole-genome sequencing with blue indicating losses and red indicating gains. Within the circle, rearrangements are shown as arcs with intrachromosomal events in green and interchromosomal translocations in purple.

# Landscape of Genomic Alterations in Cervical Carcinomas

*Ojesina AI, Lichtenstein L, et al.*

# SUPPLEMENTARY TABLES

## Supplementary Table 1. Comparison of Histological Characteristics by Geographical Source of Tumor Tissue

| | Norway<br>n=100 | Mexico<br>n=15 | Total<br>n=115 | $\chi 2$ p value |
|---|---|---|---|---|
| **Histological diagnosis** | | | | **0.3794** |
| Squamous cell carcinoma | **66 (66%)** | **13 (87%)** | **79 (69%)** | |
| Adenocarcinoma | **22 (22%)** | **2 (13%)** | **24 (21%)** | |
| Adenosquamous carcinoma | **7 (7%)** | **0 (0%)** | **7 (6%)** | |
| Neuroendocrine carcinoma | **2 (2%)** | **0 (0%)** | **2 (2%)** | |
| Clear cell carcinoma | **2 (2%)** | **0 (0%)** | **2 (2%)** | |
| Serous carcinoma | **1 (1%)** | **0 (0%)** | **1 (1%)** | |
| **Grade** | | | | **0.6712** |
| 1 | **5 (5%)** | **1 (7%)** | **6 (5%)** | |
| 2 | **45 (45%)** | **9 (60%)** | **54 (47%)** | |
| 3 | **34 (34%)** | **3 (20%)** | **37 (31%)** | |
| NA | **16 (16%)** | **2 (13%)** | **18 (16%)** | |

***The p values were obtained from 6 X 2 and 4 X 2 chi-squared analyses for histology and grade, respectively.***

**Supplementary table 2. Overview of patient characteristics for the patient cohort investigated by deep sequencing compared to the rest of the patients from the same region of Western Norway treated at Haukeland University Hospital, Bergen, Hordaland, Norway from 2001-2011.**

| Characteristics | Not characterized by deep sequencing n=249 (71%) | Characterized by deep sequencing n=100 (29%) | P-value |
|---|---|---|---|
| Age at diagnosis (mean) | 47.4 | 47.0 | 0.8* |
| Smoking habits | | | 0.7 |
| Current smoker | 95 (38%) | 40 (42%) | |
| Current non-smoker | 153 (62%) | 58 (58%) | |
| Menopausal status | | | 0.3 |
| Premenopausal | 132 (53%) | 48 (48%) | |
| Perimenopausal | 30 (12%) | 18 (18%) | |
| Postmenopausal | 87 (35%) | 34 (34%) | |
| FIGO classification | | | <0.001 |
| Stage IA | 49(20%) | 0 | |
| Stage IB | 145 (58%) | 78 (78%) | |
| Stage II | 37 (15%) | 18 (18%) | |
| Stage III | 12 (5%) | 2 (2%) | |
| Stage IV | 6 (2%) | 2 (2%) | |
| Tumor size at diagnosis | | | 0.7 |
| ≤ 4 cm | 95 (70%) | 57 (66%) | |
| > 4 cm | 41 (30%) | 29 (34%) | |
| Histological subtype | | | 0.05 |
| Adenocarcinoma | 64 (26%) | 22 (22%) | |
| Squamous cell carcinoma | 172 (69%) | 66 (66%) | |
| Other | 13 (5%) | 14 (13%) | |
| Histological differentiation* | | | 0.008 |
| Grade 1 | 33 (20%) | 5 (6%) | |
| Grade 2 | 83 (51%) | 45 (54%) | |
| Grade 3 | 47 (29%) | 34 (41%) | |
| Primary treatment | | | 0.007 |
| Surgical treatment only | 145 (58%) | 49 (49%) | |
| Surgery with adjuvant therapy | 43 (17%) | 33 (33%) | |
| Radiotherapy/chemotherapy | 59 (24%) | 16 (16%) | |
| Other | 2 (1%) | 2 (2%) | |
| % 5-year disease specific survival | 79.8% | 65.6% | 0.05** |

Data available for smoking habits (346 patients), grade (247 patients), tumor size at preoperative clinical assessment (222 patients).

P-values are based on Pearson's chi-square test, *the t-test and **the Log Rank test.

## Supplementary Table 3.  Summary Data Table

**This is an uploaded Excel Data File**

## Supplementary Table 4. Mutation types and rates across 115 cervical carcinomas

| Type of mutation | Count |
|---|---|
| Frame_Shift_Del | 292 |
| Frame_Shift_Ins | 112 |
| In_Frame_Del | 109 |
| In_Frame_Ins | 19 |
| Missense_Mutation | 11419 |
| Nonsense_Mutation | 936 |
| Nonstop_Mutation | 17 |
| Silent | 4643 |
| Splice_Site | 219 |
| Translation_Start_Site | 29 |
| **Total** | **17795** |

| | # samples | Nonsilent mutation rate (/Mb) | Nonsilent mutation rate significance | Major mutational signature | Mutation rate for major mutational signature (/Mb) |
|---|---|---|---|---|---|
| **CERVICAL CARCINOMA** | | | | | |
| All Samples | 115 | 3.689 | | Tp*C->(T/G) | 15.26 |
| Squamous cell carcinoma | 79 | 4.22 | p = 0.00949 | Tp*C->(T/G) | 18.14 |
| Adenocarcinoma | 24 | 1.605 | | *CpG->T | 8.02 |
| **OTHER CANCERS** | | | | | |
| Pediatric Rhabdoid | 35 | 0.19 | Lee, R.S. et al. J Clin Invest 122, 2983-8 (2012) | | |
| Breast Cancer | 103 | 1.27 | Banerji, S. et al. Nature 486, 405-9 (2012) | | |
| Prostate Cancer | 112 | 1.4 | Berger, M.F. et al. Nature 470, 214-20 (2011) | | |
| Head and Neck SCC | 74 | 3.3 | Stransky, N. et al. Science 333, 1157-60 (2011) | | |
| Lung Adenocarcinoma | 183 | 12 | Imielinski et al. Cell 150, 1107-20 (2012) | | |

Tp*C->(T/G): a point mutation from cytosine to thymine or guanine when the cytosine is preceded by a thymidine
*CpG->T: a point mutation from cytosine to thymidine where the cytosine precedes a guanine

## Supplementary Table 5. Relationships between epidemiological factors and mutation rates

***(See Supplementary Note 8 for discussion)***

### STRATIFICATION OF HISTOLOGY BY EPIDEMIOLOGICAL FACTORS

| Factor | Adeno | | | Squamous | | | p | Test |
|---|---|---|---|---|---|---|---|---|
| **Age** | Median: 41 years | | | Median: 47.8 years | | | 0.043 | Wilcoxon |
| **Grade** | **I** | **II** | **III** | **I** | **II** | **III** | | |
| *N* | 5 | 8 | 5 | 1 | 45 | 27 | 0.002 | Fisher Exact Test |
| *N* (Grade I removed) | 0 | 8 | 5 | 0 | 45 | 27 | 0.590 | Fisher Exact Test |
| **Geography** | **Norway** | **Mexico** | | **Norway** | **Mexico** | | | |
| *N* | 22 | 2 | | 66 | 13 | | 0.265 | Fisher Exact Test |
| **Smoking Status** | **Yes** | **No** | **NA** | **Yes** | **No** | **NA** | | |
| *N* | 7 | 17 | 0 | 35 | 42 | 2 | 0.312 | Fisher Exact Test |
| *N* (NA removed) | 7 | 17 | 0 | 35 | 42 | 0 | 0.119 | Fisher Exact Test |

### MUTATION RATE ANALYSIS

| Factor | $\log_{10}(m+1)$ median by category | | | p | Test |
|---|---|---|---|---|---|
| **Age** | | | | 0.005 | Pearson Correlation |
| **Grade** | **I** | **II** | **III** | | |
| | 0.443 | 0.458 | 0.428 | 0.679 | Kruskal-Wallis |
| **Geography** | **Norway** | **Mexico** | | | |
| | 0.413 | 0.601 | | 0.037 | Wilcoxon |
| **Smoking** | **Yes** | **No** | | | |
| | 0.459 | 0.414 | | 0.467 | Wilcoxon |
| **Histology** | **Adeno** | **Squamous** | | | |
| | 0.368 | 0.462 | | 0.009 | Wilcoxon |

m is the nonsilent mutation rate (mutations/Mb)

### LINEAR REGRESSION MODEL

$$\log_{10}(m+1) = c + B_0*x_1 + B_1*x_2$$

| Coefficient | Estimated Value | 95% CI | p |
|---|---|---|---|
| **c** | 0.259 | 0.026 to 0.493 | 0.030 |
| **$B_0$** | -0.145 | -0.287 to -0.003 | 0.045 |
| **$B_1$** | 0.006 | 0.001 to 0.011 | 0.012 |

m is the nonsilent mutation rate (mutations/Mb)

$x_1$ is histology (adeno = 1, squamous = 0)
$x_2$ is age (continuous)

## Supplementary Table 6. Validation of mutations

| Sample | GENE | Protein_Change | Powered to validate by at least one method | Validated by at least one method | RNASeq done | Adequate RNASeq Coverage | Validated by RNASeq | MiSeq done | Adequate MiSeq Coverage | Validated by MiSeq |
|---|---|---|---|---|---|---|---|---|---|---|
| SGCX-NOR-053 | B2M | p.Y46fs | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-074 | B2M | p.R3C | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-093 | B2M | p.L15F | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-003 | DDX3X | p.G635fs | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-016 | DDX3X | p.Q417H | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-045 | DDX3X | p.P568S | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-048 | DDX3X | p.M1I | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-063 | DDX3X | p.R534C | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |
| SGCX-NOR-092 | DDX3X | p.R296G | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-003 | EP300 | p.S255* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-007 | EP300 | p.E1263Q | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| SGCX-NOR-034 | EP300 | p.W1466C | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-043 | EP300 | p.Q458* | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| SGCX-NOR-045 | EP300 | p.Q2321fs | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-055 | EP300 | p.D1107H | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-055 | EP300 | p.E1514K | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-078 | EP300 | p.V1594V | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-078 | EP300 | p.A1605fs | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| SGCX-NOR-082 | EP300 | p.C1204R | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| SGCX-NOR-086 | EP300 | p.D1091H | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-094 | EP300 | p.G2196A | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-102 | EP300 | p.P1440R | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| SGCX-NOR-104 | EP300 | p.Q1489* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-005 | ERBB2 | p.D769Y | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-075 | ERBB2 | p.S310Y | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-094 | ERBB2 | p.S310F | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-097 | ERBB2 | p.V842I | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-097 | ERBB2 | p.E1067Q | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-MEX-003 | FBXW7 | p.S678* | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-MEX-015 | FBXW7 | p.R465C | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-002 | FBXW7 | p.R465H | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| SGCX-NOR-023 | FBXW7 | p.R505G | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-028 | FBXW7 | p.Q631* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-044 | FBXW7 | p.R543T | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| SGCX-NOR-045 | FBXW7 | p.R465C | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-054 | FBXW7 | p.R465C | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-077 | FBXW7 | p.R479P | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| SGCX-NOR-093 | FBXW7 | p.R505G | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| SGCX-NOR-021 | HLA-B | p.P307fs | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-023 | HLA-B | p.S35F | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| SGCX-NOR-045 | HLA-B | p.N110N | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| SGCX-NOR-057 | HLA-B | p.E299_splice | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-060 | HLA-B | p.P209fs | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-077 | HLA-B | p.E152fs | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

## Supplementary Table 6. Validation of mutations (contd.)

| Sample | GENE | Protein_Change | Powered to validate by at least one method | Validated by at least one method | RNASeq done | Adequate RNASeq Coverage | Validated by RNASeq | MiSeq done | Adequate MiSeq Coverage | Validated by MiSeq |
|---|---|---|---|---|---|---|---|---|---|---|
| SGCX-NOR-017 | KRAS | p.G13D | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-014 | MAPK1 | p.E322K | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-083 | MAPK1 | p.E322K | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-093 | MAPK1 | p.E322K | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| SGCX-NOR-053 | MAPK1 | p.E322K | Yes | No | No | Na | Na | Yes | Yes | No |
| SGCX-NOR-006 | MLL3 | p.F4382fs | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-043 | MLL3 | p.L4575L | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-045 | MLL3 | p.Q2985* | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-051 | MLL3 | p.E1436K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-078 | MLL3 | p.I1344fs | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-MEX-006 | NF2 | p.I131I | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-038 | NF2 | p.M39_splice | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-052 | NF2 | p.E107* | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| SGCX-NOR-063 | NF2 | p.R196* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-068 | NF2 | p.E443* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-001 | NFE2L2 | p.W24C | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-024 | NFE2L2 | p.R34Q | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-025 | NFE2L2 | p.E82D | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-056 | NFE2L2 | p.R34P | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-011 | PIK3CA | p.E542K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-017 | PIK3CA | p.E542K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-025 | PIK3CA | p.H1047R | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-070 | PIK3CA | p.E970K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-078 | PIK3CA | p.E542K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-086 | PIK3CA | p.E545K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-094 | PIK3CA | p.E542K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-097 | PIK3CA | p.E545K | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-102 | PIK3CA | p.H1047R | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-005 | PTEN | p.42_43insG | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |
| SGCX-NOR-025 | PTEN | p.R130G | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-025 | PTEN | p.G156W | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-034 | PTEN | p.R335* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-045 | PTEN | p.R130* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-051 | PTEN | p.R130Q | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-057 | PTEN | p.Q298* | Yes | Yes | No | Na | Na | Yes | Yes | Yes |
| SGCX-NOR-085 | PTEN | p.E307fs | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-094 | PTEN | p.H118Y | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SGCX-NOR-028 | RB1 | p.R830_splice | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-003 | STK11 | p.S216F | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-046 | STK11 | p.Y60* | Yes | Yes | Yes | Yes | Yes | No | Na | Na |
| SGCX-NOR-056 | TP53 | p.S183* | Yes | Yes | Yes | Yes | Yes | No | Na | Na |

## Supplementary Table 7. Top 50 Genes on List of Significant Mutated Genes Across 79 Squamous Cell Carcinomas of the Cervix

| Rank | Gene | Description | N | n | npat | nsite | nsil | n1 | n2 | n3 | n4 | n5 | n6 | q | $log_2$(FPKM) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **FBXW7** | **F-box and WD repeat domain containing 7** | **194988** | **12** | **12** | **8** | **0** | **3** | **7** | **0** | **0** | **2** | **0** | **4.03E-12** | **2.383** |
| **2** | **PIK3CA** | **phosphoinositide-3-kinase, catalytic, alpha polypeptide** | **258620** | **11** | **10** | **5** | **0** | **10** | **0** | **0** | **0** | **1** | **0** | **<9.08e-12** | **2.729** |
| **3** | **MAPK1** | **mitogen-activated protein kinase 1** | **78741** | **6** | **6** | **3** | **0** | **6** | **0** | **0** | **0** | **0** | **0** | **0.000671** | **3.947** |
| **4** | **HLA-B** | **major histocompatibility complex, class I, B** | **81832** | **7** | **6** | **7** | **1** | **3** | **0** | **0** | **1** | **3** | **0** | **0.00169** | **10.124** |
| **5** | **STK11** | **serine/threonine kinase 11** | **71420** | **3** | **2** | **2** | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0.012** | **3.937** |
| **6** | **EP300** | **E1A binding protein p300** | **580297** | **13** | **12** | **13** | **1** | **7** | **1** | **0** | **1** | **4** | **0** | **0.0354** | **2.678** |
| **7** | **NFE2L2** | **nuclear factor (erythroid-derived 2)-like 2** | **141629** | **3** | **3** | **2** | **0** | **3** | **0** | **0** | **0** | **0** | **0** | **0.0597** | **5.944** |
| **8** | **PTEN** | **phosphatase and tensin homolog (mutated in multiple advanced cancers 1)** | **95552** | **5** | **5** | **5** | **0** | **2** | **0** | **0** | **0** | **3** | **0** | **0.0693** | **4.041** |
| 9 | CASP8 | caspase 8, apoptosis-related cysteine peptidase | 138022 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0.11 | 3.882 |
| 10 | TP53 | tumor protein p53 | 97022 | 4 | 4 | 4 | 0 | 1 | 0 | 0 | 1 | 2 | 0 | 0.114 | 4.587 |
| 11 | HLA-A | major histocompatibility complex, class I, A | 87959 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0.281 | 9.246 |
| 12 | RB1 | retinoblastoma 1 (including osteosarcoma) | 204741 | 5 | 5 | 5 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 0.283 | 4.202 |
| 13 | POU4F2 | POU class 4 homeobox 2 | 85224 | 4 | 3 | 3 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0.283 | -6.644 |
| 14 | B2M | beta-2-microglobulin | 29370 | 3 | 3 | 3 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0.334 | 12.041 |
| 15 | SRP19 | signal recognition particle 19kDa | 35229 | 3 | 3 | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0.421 | 5.841 |
| 16 | USP9X | ubiquitin specific peptidase 9, X-linked | 616574 | 2 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0.431 | 3.889 |
| 17 | EXOC8 | exocyst complex component 8 | 172330 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0.431 | 1.887 |
| 18 | MFI2 | antigen p97 (melanoma associated) identified by monoclonal antibodies 133.2 and 96.5 | 175753 | 5 | 5 | 5 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 0.431 | 2.723 |
| 19 | PPIG | peptidylprolyl isomerase G (cyclophilin G) | 180068 | 3 | 3 | 3 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0.496 | 4.252 |
| 20 | CD68 | CD68 molecule | 84929 | 3 | 3 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0.496 | 5.151 |
| 21 | SI | sucrase-isomaltase (alpha-glucosidase) | 441822 | 6 | 6 | 6 | 1 | 3 | 0 | 0 | 3 | 0 | 0 | 0.656 | -6.644 |
| 22 | NR2E1 | nuclear receptor subfamily 2, group E, member 1 | 89233 | 2 | 2 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0.656 | -5.402 |
| 23 | PPIP5K1 | diphosphoinositol pentakisphosphate kinase 1 | 110807 | 2 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0.711 | -6.644 |
| 24 | TSG101 | tumor susceptibility gene 101 | 95035 | 3 | 3 | 3 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0.74 | 5.616 |
| 25 | ACRC | acidic repeat containing | 155437 | 2 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0.789 | -0.633 |
| 26 | ALG2 | asparagine-linked glycosylation 2 homolog (S. cerevisiae, alpha-1,3-mannosyltransferase) | 88763 | 3 | 3 | 3 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0.802 | 4.089 |
| 27 | TUBGCP6 | tubulin, gamma complex associated protein 6 | 400832 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0.802 | 4.329 |
| 28 | SLC41A1 | solute carrier family 41, member 1 | 124858 | 3 | 3 | 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0.802 | 3.022 |
| 29 | CCDC8 | coiled-coil domain containing 8 | 127706 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0.802 | 0.230 |
| 30 | HCLS1 | hematopoietic cell-specific Lyn substrate 1 | 118291 | 4 | 4 | 4 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0.802 | 4.192 |
| 31 | C16orf57 | chromosome 16 open reading frame 57 | 62213 | 2 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0.802 | 3.946 |
| 32 | APOBEC3G | apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G | 92277 | 4 | 3 | 4 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0.802 | 3.012 |

## Supplementary Table 7. Top 50 Genes on List of Significant Mutated Genes Across 79 Squamous Cell Carcinomas of the Cervix (contd.)

| Rank | Gene | Description | N | n | npat | nsite | nsil | n1 | n2 | n3 | n4 | n5 | n6 | q | $\log_2$(FPKM) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | KYNU | kynureninase (L-kynurenine hydrolase) | 115131 | 3 | 3 | 3 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0.802 | 3.170 |
| 34 | IL32 | interleukin 32 | 41510 | 3 | 3 | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0.802 | 6.921 |
| 35 | MC4R | melanocortin 4 receptor | 79237 | 2 | 2 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0.802 | -6.644 |
| 36 | NF2 | neurofibromin 2 (merlin) | 131647 | 4 | 4 | 4 | 1 | 1 | 0 | 0 | 0 | 3 | 0 | 0.802 | 2.691 |
| 37 | TRAPPC9 | trafficking protein particle complex 9 | 263311 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0.802 | 3.498 |
| 38 | SAP30BP | SAP30 binding protein | 76578 | 3 | 2 | 3 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0.877 | 4.961 |
| 39 | PIN4 | protein (peptidylprolyl cis/trans isomerase) NIMA-interacting, 4 (parvulin) | 37293 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0.979 | 3.942 |
| 40 | BAGE2 | B melanoma antigen family, member 2 | 27334 | 3 | 3 | 1 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0.979 | -6.572 |
| 41 | C2orf16 | chromosome 2 open reading frame 16 | 470720 | 4 | 2 | 4 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0.984 | -2.816 |
| 42 | IQCB1 | IQ motif containing B1 | 144667 | 3 | 3 | 3 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0.984 | 4.003 |
| 43 | LRP2 | low density lipoprotein-related protein 2 | 1121513 | 9 | 9 | 9 | 1 | 3 | 2 | 1 | 0 | 3 | 0 | 1 | -4.253 |
| 44 | LEPROTL1 | leptin receptor overlapping transcript-like 1 | 32995 | 3 | 1 | 3 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 4.087 |
| 45 | OR6K6 | olfactory receptor, family 6, subfamily K, member 6 | 81840 | 3 | 3 | 3 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | -6.644 |
| 46 | SCRIB | scribbled homolog (Drosophila) | 303759 | 2 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 5.317 |
| 47 | PRKD1 | protein kinase D1 | 205315 | 4 | 4 | 4 | 1 | 0 | 1 | 0 | 1 | 2 | 0 | 1 | -0.184 |
| 48 | ZFP36 | zinc finger protein 36, C3H type, homolog (mouse) | 78067 | 5 | 3 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 6.695 |
| 49 | ROD1 | ROD1 regulator of differentiation 1 (S. pombe) | 136438 | 2 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | N/A |
| 50 | USP28 | ubiquitin specific peptidase 28 | 258506 | 5 | 5 | 5 | 0 | 1 | 2 | 0 | 0 | 2 | 0 | 1 | 3.246 |

**Genes with q<0.1 are highlighted in bold print, but the names of non-expressed genes (FPKM < 1) are grayed out**

**N** = number of sequenced bases in this gene across the individual set
**n** = number of (nonsilent) mutations in this gene across the individual set
**npat** = number of patients (individuals) with at least one nonsilent mutation
**nsite** = number of unique sites having a nonsilent mutation
**nsil** = number of silent mutations in this gene across the individual set
**n1** = number of nonsilent mutations of type "Tp*C->(T/G)"
**n2** = number of nonsilent mutations of type "(A/C/G)p*C->(T/G)"
**n3** = number of nonsilent mutations of type "C->A"
**n4** = number of nonsilent mutations of type "A->mut"
**n5** = number of nonsilent mutations of type "indel+null"
**n6** = number of nonsilent mutations of type "double_null"
**null** = mutation category that includes nonsense, frameshift, splice-site mutations
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure)

**FPKM** = medium gene expression value in units of fragment per kilobase of exon per million fragments mapped, derived from the 79 cervical carcinoma tumors with RNASeq data

## Supplementary Table 8. Top 25 Genes on List of Significant Mutated Genes Across 24 Adenocarcinomas of the Cervix

| rank | Gene | Description | N | n | npat | nsite | nsil | n1 | n2 | n3 | n4 | n5 | n6 | q | $\log_2$(FPKM) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **ELF3** | **E74-like factor 3 (ets domain transcription factor, epithelial-specific )** | **27412** | **3** | **3** | **3** | **0** | **0** | **0** | **0** | **0** | **3** | **0** | **0.03** | **8.175** |
| **2** | **CBFB** | **core-binding factor, beta subunit** | **13943** | **2** | **2** | **2** | **0** | **0** | **0** | **0** | **1** | **1** | **0** | **0.0342** | **4.267** |
| 3 | KRAS | v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog | 16711 | 2 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0.131 | **3.125** |
| 4 | C9orf71 | chromosome 9 open reading frame 71 | 12502 | 3 | 2 | 2 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0.131 | -0.511 |
| 5 | CDH19 | cadherin 19, type 2 | 55589 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0.183 | -4.556 |
| 6 | PIK3CA | phosphoinositide-3-kinase, catalytic, alpha polypeptide | 78648 | 4 | 3 | 3 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0.329 | 2.698 |
| 7 | CTNNBL1 | catenin, beta like 1 | 39978 | 2 | 2 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.481 | 4.900 |
| 8 | FAM55A | family with sequence similarity 55, member A | 29607 | 2 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0.6 | -6.474 |
| 9 | CD83 | CD83 molecule | 14282 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.747 | 2.079 |
| 10 | PHOSPHO1 | phosphatase, orphan 1 | 8421 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.747 | N/A |
| 11 | AHNAK | AHNAK nucleoprotein | 426884 | 2 | 1 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0.769 | 6.130 |
| 12 | ZFHX4 | zinc finger homeobox 4 | 229544 | 2 | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0.769 | -1.566 |
| 13 | ELOVL4 | elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 4 | 23160 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0.858 | -4.896 |
| 14 | PLXNB3 | plexin B3 | 121403 | 3 | 3 | 3 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0.938 | 3.202 |
| 15 | ARID1A | AT rich interactive domain 1A (SWI-like) | 142849 | 3 | 3 | 3 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0.938 | 4.073 |
| 16 | DNAJB8 | DnaJ (Hsp40) homolog, subfamily B, member 8 | 16872 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.938 | -6.644 |
| 17 | HDAC8 | histone deacetylase 8 | 29643 | 2 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0.938 | 3.927 |
| 18 | ZNF268 | zinc finger protein 268 | 7640 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.938 | 1.661 |
| 19 | MIA3 | melanoma inhibitory activity family, member 3 | 136266 | 3 | 3 | 3 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0.938 | 4.917 |
| 20 | RABAC1 | Rab acceptor 1 (prenylated) | 12528 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.938 | 5.873 |
| 21 | ERGIC2 | ERGIC and golgi 2 | 27376 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.938 | 4.199 |
| 22 | NARF | nuclear prelamin A recognition factor | 34226 | 2 | 2 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0.944 | 4.491 |
| 22 | OR8H1 | olfactory receptor, family 8, subfamily H, member 1 | 22512 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.944 | -6.644 |
| 24 | SCAI | suppressor of cancer cell invasion | 45788 | 2 | 2 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.939 | 0.269 |
| 25 | CCNA2 | cyclin A2 | 31043 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.943 | 3.666 |

**Genes with q<0.1 are highlighted in bold print, but the names of non-expressed genes (FPKM < 1) are grayed out**

**N** = number of sequenced bases in this gene across the individual set
**n** = number of (nonsilent) mutations in this gene across the individual set
**npat** = number of patients (individuals) with at least one nonsilent mutation
**nsite** = number of unique sites having a nonsilent mutation
**nsil** = number of silent mutations in this gene across the individual set
**n1** = number of nonsilent mutations of type "*CpG->T "
**n2** = number of nonsilent mutations of type "*Cp(A/C/T)->T"
**n3** = number of nonsilent mutations of type "C->(G/A)"
**n4** = number of nonsilent mutations of type "A->mut"
**n5** = number of nonsilent mutations of type "indel+null"
**n6** = number of nonsilent mutations of type "double_null"
**null** = mutation category that includes nonsense, frameshift, splice-site mutations
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure)

**FPKM** = medium gene expression value in units of fragment per kilobase of exon per million fragments mapped, derived from the 79cervical carcinoma tumors with RNASeq data

## Supplementary Table 9a. Significant Mutated Genes in COSMIC Territory Across 79 Squamous Cell Carcinomas of the Cervix with q<0.1

| Rank | Gene | Description | n | cos | n_cos | N_cos | cos_ev | q |
|---|---|---|---|---|---|---|---|---|
| **1** | **FBXW7** | **F-box and WD repeat domain containing 7** | **12** | **91** | **10** | **7189** | **495** | **1.07E-09** |
| **2** | **PIK3CA** | **phosphoinositide-3-kinase, catalytic, alpha polypeptide** | **11** | **184** | **9** | **14536** | **3955** | **1.07E-09** |
| 3 | TP53 | tumor protein p53 | 4 | 308 | 4 | 24332 | 211 | 7.54E-03 |
| 4 | PTEN | phosphatase and tensin homolog (mutated in multiple advanced cancers 1) | 5 | 728 | 5 | 57512 | 404 | 7.87E-03 |
| 5 | STK11 | serine/threonine kinase 11 | 3 | 130 | 3 | 10270 | 8 | 1.34E-02 |
| **6** | **ERBB2** | **v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)** | **4** | **41** | **2** | **3239** | **4** | **7.56E-02** |

**Genes with novel mutations previously unreported in COSMIC are highlighted in bold font , with difference in mutation count highlighted in red.**

**n** - number of (nonsilent) mutations in this gene across the individual set.
**cos** = number of unique mutated sites in this gene in COSMIC
**n_cos** = overlap between n and cos.
**N_cos** = number of individuals times cos.
**cos_ev** = total evidence: number of reports in COSMIC for mutations seen in this gene.
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure) of sequenced bases in this gene across the 81 samples

The previously unreported mutations in squamous cell carcinomas are:

**FBXW7: Q631*, R678***

**PIK3CA: L267V, R577K**

**ERBB2: P579R, E1067Q** *(one adenocarcinoma sample has an R678Q mutation in ERBB2 as well)*

# Supplementary Table 9b. Significant Mutated Genes in COSMIC Territory Across 24 Adenocarcinomas of the Cervix with q<0.1

| Rank | Gene | Description | n | cos | n_cos | N_cos | cos_ev | q |
|---|---|---|---|---|---|---|---|---|
| 1 | PIK3CA | phosphoinositide-3-kinase, catalytic, alpha polypeptide | 4 | 184 | 4 | 4416 | 682 | 6.14E-07 |
| 2 | KRAS | v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog | 2 | 51 | 2 | 1224 | 16622 | 4.97E-03 |
| 3 | SCAI | suppressor of cancer cell invasion | 2 | 1 | 1 | 24 | 1 | 4.65E-02 |
| 4 | SNX25 | sorting nexin 25 | 1 | 1 | 1 | 24 | 1 | 4.65E-02 |
| **5** | **TTK** | **TTK protein kinase** | **2** | **2** | **1** | **48** | **3** | **7.45E-02** |

**Genes with novel mutations previously unreported in COSMIC are highlighted in bold font , with difference in mutation count highlighted in red.**

**n** - number of (nonsilent) mutations in this gene across the individual set.
**cos** = number of unique mutated sites in this gene in COSMIC
**n_cos** = overlap between n and cos.
**N_cos** = number of individuals times cos.
**cos_ev** = total evidence: number of reports in COSMIC for mutations seen in this gene.
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure) of sequenced bases in this gene across the 81 samples

## Supplementary Table 10a. Significantly Mutated Gene Sets in 79 Squamous Carcinomas of the Cervix

| Geneset | Mutated genes within geneset | q1 | q2 |
|---|---|---|---|
| **TIDPATHWAY** | **IFNG(1), IFNGR1(4), IKBKB(1), JAK2(4), NFKBIA(1), RB1(5), TNFRSF1A(1), TP53(4), WT1(2)** | **1.50E-05** | **2.88E-05** |
| **HSA00232_CAFFEINE_METABOLISM** | **CYP2A13(2), CYP2A6(2), CYP2A7(4), NAT2(1), XDH(4)** | **0.00677** | **0.0169** |
| **IFNGPATHWAY** | **IFNG(1), IFNGR1(4), JAK1(1), JAK2(4), STAT1(1)** | **0.027** | **0.0809** |
| HCMVPATHWAY | AKT1(2), CREB1(1), MAP2K2(1), MAP2K3(1), MAP2K6(2), MAP3K1(1), MAPK1(6), MAPK14(1), PIK3CA(11), RB1(5), SP1(2) | 1.50E-05 | 0.0831 |
| **TERCPATHWAY** | **NFYC(1), RB1(5), SP1(2)** | **0.0415** | **0.0831** |
| TERTPATHWAY | HDAC1(1), SP1(2), TP53(4), WT1(2) | 0.0585 | 0.107 |
| ST_INTERFERON_GAMMA_PATHWAY | IFNG(1), IFNGR1(4), JAK1(1), JAK2(4), PLA2G2A(1), PTPRU(2), STAT1(1) | 0.0731 | 0.136 |
| CTLPATHWAY | B2M(3), CD3E(1), HLA-A(4), ICAM1(1), ITGAL(1), PRF1(1) | 0.0731 | 0.136 |
| ARFPATHWAY | ABL1(3), CDKN2A(2), PIK3CA(11), POLR1A(2), POLR1B(1), RAC1(1), RB1(5), TP53(4) | 0.00351 | 0.188 |
| NKCELLSPATHWAY | B2M(3), HLA-A(4), ITGB1(2), KLRC1(1), KLRC3(2), PIK3CA(11), PTPN6(3), RAC1(1) | 0.0105 | 0.491 |
| FBW7PATHWAY | CDC34(1), CUL1(1), FBXW7(12), RB1(5) | 8.23E-05 | 1 |
| CTLA4PATHWAY | CD28(2), CD3E(1), CTLA4(1), GRB2(1), ICOS(1), IL2(1), ITK(2), PIK3CA(11) | 0.0202 | 1 |
| PTENPATHWAY | AKT1(2), BCAR1(1), GRB2(1), ITGB1(2), MAPK1(6), PDPK1(1), PIK3CA(11), PTEN(5), PTK2(1), SOS1(3) | 0.0202 | 1 |
| RACCYCDPATHWAY | AKT1(2), MAPK1(6), NFKBIA(1), PIK3CA(11), RAC1(1), RAF1(1), RB1(5) | 0.0731 | 1 |
| PELP1PATHWAY | CREBBP(4), EP300(13), MAPK1(6) | 0.0731 | 1 |

Recurrently mutated genes (q<0.1) are highlighted in red.
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure)

q1 = obtained from analysis of all genes in each geneset
q2 = obtained from analysis that excluded the statistically recurrently mutated genes highlighted in red

## Supplementary Table 10b. Significantly Mutated Gene Sets in 24 Adenocarcinomas of the Cervix

| Geneset | Mutated genes within geneset | q1 | q2 |
|---|---|---|---|
| PTENPATHWAY | AKT1(1), PIK3CA(4), PTEN(2), SOS1(1) | 0.0143 | 0.0143 |
| SA_PTEN_PATHWAY | AKT1(1), PIK3CA(4), PTEN(2), SOS1(1) | 0.0143 | 0.0143 |
| IGF1MTORPATHWAY | AKT1(1), PIK3CA(4), PTEN(2) | 0.037 | 0.037 |
| SA_TRKA_RECEPTOR | AKT1(1), CDKN1A(1), PIK3CA(4), SOS1(1) | 0.037 | 0.037 |
| GSK3PATHWAY | AKT1(1), APC(2), IRAK1(1), PIK3CA(4) | 0.0376 | 0.0376 |
| TRKAPATHWAY | AKT1(1), PIK3CA(4), SOS1(1) | 0.0815 | 0.0815 |
| ACHPATHWAY | AKT1(1), PIK3CA(4), RAPSN(1) | 0.0815 | 0.0815 |
| HCMVPATHWAY | AKT1(1), MAP2K3(1), PIK3CA(4) | 0.0815 | 0.0815 |
| MTORPATHWAY | AKT1(1), PIK3CA(4), PTEN(2) | 0.0815 | 0.0815 |
| ST_PHOSPHOINOSITIDE_3_KINASE_PATHWAY | AKT1(1), PIK3CA(4), PTEN(2), RPS6KA3(1), SOS1(1), YWHAB(1) | 0.0815 | 0.0815 |
| EIF4PATHWAY | AKT1(1), PIK3CA(4), PTEN(2) | 0.0815 | 0.0815 |
| SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES | AKT1(1), BRD4(1), PIK3CA(4), PTEN(2), RPS6KA3(1), SERPINB6(1), SOS1(1), YWHAB(1) | 0.0982 | 0.0982 |
| GCRPATHWAY | AKT1(1), GNAS(2), PIK3CA(4) | 0.0982 | 0.0982 |
| PLCPATHWAY | AKT1(1), PIK3CA(4) | 0.0982 | 0.0982 |
| IGF1RPATHWAY | AKT1(1), PIK3CA(4), SOS1(1) | 0.0982 | 0.0982 |

Recurrently mutated genes ($q<0.1$) are highlighted in red (in this case, none)
**q** = q-value, False Discovery Rate (Benjamini-Hochberg procedure)

q1 = obtained from analysis of all genes in each geneset
q2 = obtained from analysis that excluded the statistically recurrently mutated genes highlighted in red (hence q1=q2)

# Supplementary Table 11. Somatic copy number alterations across 79 squamous cell carcinomas of the cervix

## Broad arm-level gains

| Arm | # Genes | Amplification Frequency | Z score | Q value | Previously reported | References |
|---|---|---|---|---|---|---|
| 3q | 1139 | 0.62 | 9.43 | 0 | yes | 1,2 |
| 1q | 1955 | 0.36 | 5.38 | 7.11E-07 | yes | 2,3 |
| 1p | 2121 | 0.33 | 4.75 | 1.32E-05 | | |
| 20p | 355 | 0.33 | 2.83 | 0.0183 | | |
| 20q | 753 | 0.31 | 2.83 | 0.0183 | yes | 1,4 |
| 14q | 1341 | 0.26 | 2.15 | 0.102 | | |
| 5p | 270 | 0.28 | 1.94 | 0.145 | | |
| 19q | 1709 | 0.23 | 1.82 | 0.169 | | |
| 8q | 859 | 0.25 | 1.58 | 0.245 | | |

## Broad arm-level losses

| Arm | # Genes | Deletion Frequency | Z score | Q value | Previously reported | References |
|---|---|---|---|---|---|---|
| 3p | 1062 | 0.51 | 7.79 | 1.34E-13 | yes | 2,3 |
| 4p | 489 | 0.47 | 6.27 | 3.51E-09 | yes | 3,5 |
| 13q | 654 | 0.41 | 5.14 | 1.78E-06 | yes | 2,3 |
| 3q | 1139 | 0.44 | 4.59 | 2.14E-05 | | |
| 4q | 1049 | 0.34 | 3.92 | 0.000347 | yes | 5,6,7 |
| 11q | 1515 | 0.32 | 3.81 | 0.000452 | yes | 1, 3, 8, 9 |
| 17p | 683 | 0.34 | 3.45 | 0.00154 | yes | 9 |
| 11p | 862 | 0.32 | 3.23 | 0.00306 | | |
| 6q | 839 | 0.29 | 2.53 | 0.0245 | yes | 3 |
| 8p | 580 | 0.27 | 1.82 | 0.134 | yes | 12 |
| 6p | 1173 | 0.24 | 1.52 | 0.226 | yes | 5,7,10,11 |

1. Wilting, S.M. et al. Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 47, 890-905 (2008).
2. Matthews, C.P., Shera, K.A. & McDougall, J.K. Genomic changes and HPV type in cervical carcinoma. Proc Soc Exp Biol Med 223, 316-21 (2000).
3. Kirchhoff, M. et al. Comparative genomic hybridization reveals a recurrent pattern of chromosomal aberrations in severe dysplasia/carcinoma in situ of the cervix and in advanced-stage cervical carcinoma. Genes Chromosomes Cancer 24, 144-50 (1999).
4. Scotto, L. et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer 47, 755-65 (2008).
5. Umayahara, K. et al. Comparative genomic hybridization detects genetic alterations during early stages of cervical cancer progression. Genes Chromosomes Cancer 33, 98-102 (2002).
6. Dellas, A., Torhorst, J., Gaudenz, R., Mihatsch, M.J. & Moch, H. DNA copy number changes in cervical adenocarcinoma. Clin Cancer Res 9, 2985-91 (2003).
7. Fitzpatrick, M.A. et al. Identification of chromosomal alterations important in the development of cervical intraepithelial neoplasia and invasive carcinoma using alignment of DNA microarray data. Gynecol Oncol 103, 458-62 (2006).
8. Huang, K.F. et al. Chromosomal gain of 3q and loss of 11q often associated with nodal metastasis in early stage cervical squamous cell carcinoma. J Formos Med Assoc 106, 894-902 (2007).
9. Kersemaekers, A.M., Hermans, J., Fleuren, G.J. & van de Vijver, M.J. Loss of heterozygosity for defined regions on chromosomes 3, 11 and 17 in carcinomas of the uterine cervix. Br J Cancer 77, 192-200 (1998).
10. Mazurenko, N. et al. High resolution mapping of chromosome 6 deletions in cervical cancer. Oncol Rep 6, 859-63 (1999).
11. Koopman, L.A., Corver, W.E., van der Slik, A.R., Giphart, M.J. & Fleuren, G.J. Multiple genetic alterations cause frequent and heterogeneous human histocompatibility leukocyte antigen class I loss in cervical cancer. J Exp Med 191, 961-76 (2000).
12. Jee, K.J., Kim, Y.T., Kim, K.R., Aalto, Y. & Knuutila, S. Amplification at 9p in cervical carcinoma by comparative genomic hybridization. Anal Cell Pathol 22, 159-63 (2001).
13. Tsuda, H. et al. Different pattern of loss of heterozygosity among endocervical-type adenocarcinoma, endometrioid-type adenocarcinoma and adenoma malignum of the uterine cervix. Int J Cancer 98, 713-7 (2002).
14. Sherwood, J.B. et al. Chromosome 4 deletions are frequent in invasive cervical cancer and differ between histologic variants. Gynecol Oncol 79, 90-6 (2000).

## SupplementaryTable 12. Somatic copy number alterations across 24 adenocarcinomas of the cervix

### Broad arm-level gains

| Arm | # Genes | Amplification Frequency | Z score | Q value | Previously reported | References |
|---|---|---|---|---|---|---|
| 1q | 1955 | 0.35 | 2.84 | 0.0439 | yes | 1 |
| 3q | 1139 | 0.39 | 2.99 | 0.0439 | yes | 1 |
| 19q | 1709 | 0.32 | 2.26 | 0.129 | yes | 13 |
| 20p | 355 | 0.36 | 2.22 | 0.129 | | |

### Broad arm-level losses

| Arm | # Genes | Deletion Frequency | Z score | Q value | Previously reported | References |
|---|---|---|---|---|---|---|
| 18q | 446 | 0.54 | 4.63 | 7.00E-05 | yes | 6 |
| 4p | 489 | 0.46 | 3.59 | 0.00304 | yes | 14 |
| 16q | 702 | 0.45 | 3.5 | 0.00304 | yes | 13 |
| 4q | 1049 | 0.42 | 3.34 | 0.00404 | yes | 6 |
| 11q | 1515 | 0.35 | 2.61 | 0.0355 | yes | 9 |
| 11p | 862 | 0.35 | 2.29 | 0.0707 | yes | 9 |
| 16p | 872 | 0.33 | 2.02 | 0.121 | | |
| 19p | 995 | 0.3 | 1.79 | 0.177 | yes | 13 |

1. Wilting, S.M. et al. Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 47, 890-905 (2008).
2. Matthews, C.P., Shera, K.A. & McDougall, J.K. Genomic changes and HPV type in cervical carcinoma. Proc Soc Exp Biol Med 223, 316-21 (2000).
3. Kirchhoff, M. et al. Comparative genomic hybridization reveals a recurrent pattern of chromosomal aberrations in severe dysplasia/carcinoma in situ of the cervix and in advanced-stage cervical carcinoma. Genes Chromosomes Cancer 24, 144-50 (1999).
4. Scotto, L. et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer 47, 755-65 (2008).
5. Umayahara, K. et al. Comparative genomic hybridization detects genetic alterations during early stages of cervical cancer progression. Genes Chromosomes Cancer 33, 98-102 (2002).
6. Dellas, A., Torhorst, J., Gaudenz, R., Mihatsch, M.J. & Moch, H. DNA copy number changes in cervical adenocarcinoma. Clin Cancer Res 9, 2985-91 (2003).
7. Fitzpatrick, M.A. et al. Identification of chromosomal alterations important in the development of cervical intraepithelial neoplasia and invasive carcinoma using alignment of DNA microarray data. Gynecol Oncol 103, 458-62 (2006).
8. Huang, K.F. et al. Chromosomal gain of 3q and loss of 11q often associated with nodal metastasis in early stage cervical squamous cell carcinoma. J Formos Med Assoc 106, 894-902 (2007).
9. Kersemaekers, A.M., Hermans, J., Fleuren, G.J. & van de Vijver, M.J. Loss of heterozygosity for defined regions on chromosomes 3, 11 and 17 in carcinomas of the uterine cervix. Br J Cancer 77, 192-200 (1998).
10. Mazurenko, N. et al. High resolution mapping of chromosome 6 deletions in cervical cancer. Oncol Rep 6, 859-63 (1999).
11. Koopman, L.A., Corver, W.E., van der Slik, A.R., Giphart, M.J. & Fleuren, G.J. Multiple genetic alterations cause frequent and heterogeneous human histocompatibility leukocyte antigen class I loss in cervical cancer. J Exp Med 191, 961-76 (2000).
12. Jee, K.J., Kim, Y.T., Kim, K.R., Aalto, Y. & Knuutila, S. Amplification at 9p in cervical carcinoma by comparative genomic hybridization. Anal Cell Pathol 22, 159-63 (2001).
13. Tsuda, H. et al. Different pattern of loss of heterozygosity among endocervical-type adenocarcinoma, endometrioid-type adenocarcinoma and adenoma malignum of the uterine cervix. Int J Cancer 98, 713-7 (2002).
14. Sherwood, J.B. et al. Chromosome 4 deletions are frequent in invasive cervical cancer and differ between histologic variants. Gynecol Oncol 79, 90-6 (2000).

**Supplementary Table 13. Correlation of somatic copy number alterations and gene expression in 16898 genes across 79 cervical carcinomas**

**This is an uploaded Excel Data File**

# Supplementary Table 14. HPV Typing

| | | HPV type | | | | |
|---|---|---|---|---|---|---|
| Sample | Histology | f-HPV | MASSArray | RNASeq | WGS | TP53 Mutation Status |
| SGCX-NOR-001 | Squamous cell carcinoma | 18,16 | 16, 18 | n/a | n/a | WT |
| SGCX-NOR-002 | Squamous cell carcinoma | 45,16 | 16, 31, 45 | 45 | 45 | WT |
| SGCX-NOR-003 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-005 | Adenocarcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-006 | Squamous cell carcinoma | 52,16 | Negative | 52 | n/a | WT |
| SGCX-NOR-007 | Squamous cell carcinoma | 16 | 16, 18 | 16 | n/a | WT |
| SGCX-NOR-008 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-010 | Squamous cell carcinoma | 16,18 | 16, 18 | n/a | n/a | WT |
| SGCX-NOR-011 | Neuroendocrine carcinoma | 18 | Negative | 18 | n/a | WT |
| SGCX-NOR-012 | Squamous cell carcinoma | 16,18 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-013 | Neuroendocrine carcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-014 | Squamous cell carcinoma | 52 | 16, 52 | 52 | n/a | WT |
| SGCX-NOR-015 | Adenocarcinoma | 16,18 | 16, 18 | 18,16 | n/a | WT |
| SGCX-NOR-016 | Adenosquamous carcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-017 | Adenocarcinoma | 16 | 16, 52 | 16 | n/a | WT |
| SGCX-NOR-018 | Squamous cell carcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-019 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-020 | Adenocarcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-021 | Adenosquamous carcinoma | 16,18 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-022 | Squamous cell carcinoma | 45,33 | Negative | n/a | n/a | WT |
| SGCX-NOR-023 | Squamous cell carcinoma | 45,16 | Negative | 45 | n/a | WT |
| SGCX-NOR-024 | Squamous cell carcinoma | 16 | 16, 18 | 16 | n/a | WT |
| SGCX-NOR-025 | Clear cell carcinoma | 16 | 16 | Negative | Negative | WT |
| SGCX-NOR-026 | Adenosquamous carcinoma | 16 | Negative | Negative | Negative | R267P |
| SGCX-NOR-027 | Squamous cell carcinoma | 16 | 16 | Negative | n/a | WT |
| SGCX-NOR-028 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-029 | Adenocarcinoma | 16 | 16, 52 | 16 | n/a | WT |
| SGCX-NOR-030 | Squamous cell carcinoma | 16 | Negative | 16 | 16 | WT |
| SGCX-NOR-031 | Squamous cell carcinoma | 33 | 16, 33 | 33 | n/a | WT |
| SGCX-NOR-032 | Adenocarcinoma | 18,16 | 16, 18, 33, 45 | Negative | n/a | WT |
| SGCX-NOR-033 | Adenocarcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-034 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-035 | Adenocarcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-036 | Adenocarcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-037 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-038 | Squamous cell carcinoma | 16,45 | 16, 45 | n/a | n/a | WT |
| SGCX-NOR-039 | Squamous cell carcinoma | 45,16 | 16, 45 | n/a | n/a | WT |
| SGCX-NOR-040 | Squamous cell carcinoma | 16,51,52 | 16, 18, 52 | n/a | n/a | WT |
| SGCX-NOR-041 | Adenocarcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-042 | Squamous cell carcinoma | 33,16 | 16, 33 | 33 | n/a | WT |
| SGCX-NOR-043 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-044 | Squamous cell carcinoma | 33,16 | 16, 31 | 31 | n/a | R196Q |
| SGCX-NOR-045 | Squamous cell carcinoma | 16,33 | 16 | 16 | n/a | WT |
| SGCX-NOR-046 | Adenocarcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-047 | Adenocarcinoma | 18 | 18 | *18* | 18 | WT |
| SGCX-NOR-048 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-049 | Serous carcinoma | 16 | Negative | n/a | n/a | WT |
| SGCX-NOR-050 | Adenosquamous carcinoma | 18 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-051 | Squamous cell carcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-052 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-053 | Squamous cell carcinoma | 16 | 16, 18 | n/a | n/a | WT |
| SGCX-NOR-054 | Squamous cell carcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-055 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-056 | Squamous cell carcinoma | 16 | 16 | Negative | Negative | S183* |
| SGCX-NOR-057 | Squamous cell carcinoma | 52 | 16, 52 | n/a | n/a | WT |
| SGCX-NOR-058 | Squamous cell carcinoma | 45,58 | 16, 45, 58 | 45 | n/a | WT |
| SGCX-NOR-059 | Squamous cell carcinoma | 16,18,33 | 16, 18 | 18 | 18 | WT |
| SGCX-NOR-060 | Squamous cell carcinoma | 16 | 16 | n/a | n/a | WT |

## Supplementary Table 14. HPV Typing (contd)

| | | HPV type | | | | |
|---|---|---|---|---|---|---|
| **Sample** | **Histology** | **f-HPV** | **MASSArray** | **RNASeq** | **WGS** | **TP53 Mutation Status** |
| SGCX-NOR-061 | Squamous cell carcinoma | 51,16 | 16, 18 | 82 | n/a | WT |
| SGCX-NOR-062 | Squamous cell carcinoma | 16 | 16 | n/a | 16 | WT |
| SGCX-NOR-063 | Adenosquamous carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-064 | Squamous cell carcinoma | 16,18 | 16, 18 | 18,16 | n/a | WT |
| SGCX-NOR-065 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-066 | Squamous cell carcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-068 | Squamous cell carcinoma | 16 | 16 | 16,18 | n/a | WT |
| SGCX-NOR-069 | Adenocarcinoma | 16 | 16 | 16,18 | n/a | WT |
| SGCX-NOR-070 | Squamous cell carcinoma | 16,56 | 16, 56 | 56,18,16 | n/a | WT |
| SGCX-NOR-071 | Squamous cell carcinoma | 16 | 16 | n/a | n/a | WT |
| SGCX-NOR-072 | Adenocarcinoma | 16,18 | 16, 18 | 18,16 | n/a | WT |
| SGCX-NOR-073 | Adenocarcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-074 | Squamous cell carcinoma | n/a | 16 | 16 | n/a | WT |
| SGCX-NOR-075 | Adenocarcinoma | 16,18 | 16, 18 | 18,16 | 18,16 | WT |
| SGCX-NOR-076 | Squamous cell carcinoma | 16,18 | 16, 18 | 16 | n/a | WT |
| SGCX-NOR-077 | Squamous cell carcinoma | 16,31 | 16, 31 | 31 | n/a | WT |
| SGCX-NOR-078 | Squamous cell carcinoma | 16,52 | 16, 52 | 52,16 | n/a | WT |
| SGCX-NOR-080 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-081 | Squamous cell carcinoma | 31,16 | 16, 31 | n/a | n/a | WT |
| SGCX-NOR-082 | Squamous cell carcinoma | 56,16 | 16, 56 | 56 | n/a | WT |
| SGCX-NOR-083 | Squamous cell carcinoma | 33,16 | 16, 33 | 33 | n/a | WT |
| SGCX-NOR-084 | Squamous cell carcinoma | 33,16 | 16 | 33 | 33 | WT |
| SGCX-NOR-085 | Adenocarcinoma | 16,33 | 16 | 16 | n/a | WT |
| SGCX-NOR-086 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-087 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-088 | Adenosquamous carcinoma | 16,18 | 16, 18 | 18 | 18 | WT |
| SGCX-NOR-089 | Adenosquamous carcinoma | 16,18 | 16, 18 | 18,16 | n/a | WT |
| SGCX-NOR-090 | Adenocarcinoma | 18 | 16, 18, 31 | 18 | n/a | WT |
| SGCX-NOR-091 | Squamous cell carcinoma | 45 | 16, 45 | 45 | n/a | WT |
| SGCX-NOR-092 | Adenocarcinoma | 16 | 16, 31 | 16 | n/a | WT |
| SGCX-NOR-093 | Squamous cell carcinoma | 16 | 16 | 16 | 16 | WT |
| SGCX-NOR-094 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-095 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-096 | Adenocarcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-097 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-098 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-099 | Squamous cell carcinoma | 16,33,18 | 16, 33 | n/a | n/a | WT |
| SGCX-NOR-100 | Squamous cell carcinoma | 16,18 | 16, 18 | n/a | n/a | V274D |
| SGCX-NOR-101 | Adenocarcinoma | 16 | 16, 45 | 16 | n/a | WT |
| SGCX-NOR-102 | Clear cell carcinoma | 18,16 | 16, 18 | 18 | n/a | WT |
| SGCX-NOR-103 | Squamous cell carcinoma | 16 | 16 | 16 | n/a | WT |
| SGCX-NOR-104 | Adenocarcinoma | 16 | 16 | 16 | 16 | WT |
| SGCX-MEX-001 | Squamous cell carcinoma | *18* | 18 | n/a | n/a | WT |
| SGCX-MEX-002 | Adenocarcinoma | Negative | Negative | n/a | n/a | R175H |
| SGCX-MEX-003 | Squamous cell carcinoma | 18 | 18 | n/a | n/a | WT |
| SGCX-MEX-004 | Squamous cell carcinoma | 68 | 68 | n/a | n/a | WT |
| SGCX-MEX-005 | Squamous cell carcinoma | *16* | 31 | n/a | n/a | WT |
| SGCX-MEX-006 | Squamous cell carcinoma | Negative | 16, 31 | n/a | n/a | WT |
| SGCX-MEX-007 | Squamous cell carcinoma | Negative | 31 | n/a | n/a | Q192* |
| SGCX-MEX-008 | Squamous cell carcinoma | 18 | 18 | n/a | n/a | WT |
| SGCX-MEX-009 | Squamous cell carcinoma | 16 | n/a | n/a | n/a | WT |
| SGCX-MEX-010 | Squamous cell carcinoma | 16 | n/a | n/a | n/a | WT |
| SGCX-MEX-011 | Squamous cell carcinoma | 45,16 | 45 | n/a | n/a | WT |
| SGCX-MEX-012 | Squamous cell carcinoma | n/a | n/a | n/a | n/a | WT |

**Supplementary Table 15a. Data on HPV integration and other HPV analyses**

**Supplementary Table 15b. Source Data for Figure 3**

**These are uploaded Excel Data Files**

# Supplementary Table 16. MSigDB Analyses of Selected GeneSets overlapping with HPV Integration Sites

| Gene Set Name | Description | # Genes | p value | HPV Integration sites overlapping with GeneSet |
|---|---|---|---|---|
| SMID_BREAST_CANCER_BASAL_DN | Genes down-regulated in basal subtype of breast cancer samples. | 7 | 1.47E-06 | *ERBB2,RARA,PTPRT,BLVRA,DACH1,CEACAM5,TNIK* |
| GRESHOCK_CANCER_COPY_NUMBER_UP | Genes from common genomic gains observed in a meta analyis of copy number alterations across a panel of different cancer cell lines and tumor samples. | 5 | 6.99E-06 | *ERBB2,MYC,RARA,BCL11B,FANCC* |
| PID_RXR_VDR_PATHWAY | RXR and RAR heterodimerization with other nuclear receptor | 2 | 2.22E-04 | *RPS6KB1,RARA* |
| DANG_BOUND_BY_MYC | Genes whose promoters are bound by MYC [GeneID=4609], according to MYC Target Gene Database. | 6 | 2.70E-04 | *RPS6KB1,RARA,MYC,ERBB2,CEACAM5,RAP2B* |
| NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON | Genes within amplicon 8q23-q24 identified in a copy number alterations study of 191 breast tumor samples. | 3 | 3.15E-04 | *MYC,MAFA,EIF2C2* |
| HEDENFALK_BREAST_CANCER_BRCA1_VS_BRCA2 | Genes differentially expressed in hereditary breast cancer tumors with mutated BRCA1 [GeneID=672] compared to those with mutated BRCA2 [GeneID=675]. | 3 | 3.52E-04 | *MYC,EIF2C2,ERBB2* |
| PID_IL2_PI3KPATHWAY | IL2 signaling events mediated by PI3K | 2 | 3.82E-04 | *MYC,RPS6KB1* |
| SMID_BREAST_CANCER_LUMINAL_B_UP | Genes up-regulated in the luminal B subtype of breast cancer. | 3 | 4.12E-04 | *CEACAM5,DACH1,PTPRT* |
| LI_AMPLIFIED_IN_LUNG_CANCER | Genes with increased copy number that correlates with increased expression across six different lung adenocarcinoma cell lines. | 3 | 4.55E-04 | *ERBB2,CEACAM5,P4HB* |
| JNK_DN.V1_DN | Genes down-regulated in JNK inhibitor-treated (SP600125[PubChem=8515]) keratinocytes. | 3 | 5.59E-04 | *EIF2C2,RARA,TNIK* |
| E2F1_UP.V1_DN | Genes down-regulated in mouse fibroblasts over-expressing E2F1 [Gene ID=1869] gene. | 3 | 5.76E-04 | *MYC,RAP2B,FMO3* |
| chr8q24 | Genes in cytogenetic band chr8q24 | 3 | 7.77E-04 | *MYC,MAFA,EIF2C2* |
| LOCKWOOD_AMPLIFIED_IN_LUNG_CANCER | Overexpressed genes with amplified copy number across 27 non-small cell lung cancer (NSCLC) cell lines. | 3 | 7.77E-04 | *EIF2C2,MYC,BCL11B* |
| AACTTT_UNKNOWN | Genes with promoter regions [-2kb,2kb] around transcription start site containing motif AACTTT. Motif does not match any known transcription factor | 7 | 7.95E-04 | *EIF2C2,MYC,BCL11B,ERBB2,RAP2B,DACH1,MAFA* |
| RTAAACA_V$FREAC2_01 | Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif RTAAACA which matches annotation for FOXF2: forkhead box F2 | 5 | 9.42E-04 | *MYC,BCL11B,ERBB2,RARA,BCL2L13* |
| V$P53_DECAMER_Q2 | Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif RGRCAWGNCY which matches annotation for TP53: tumor protein p53 (Li-Fraumeni syndrome) | 3 | 1.30E-03 | *BCL11B,RARA,RAP2B* |
| BRUINS_UVC_RESPONSE_VIA_TP53_GROUP_D | Category D genes: p53-independent genes whose expression in the absence of S389 phosphorylation is similar to loss of TP53 [GeneID=7157] in MEF (embryonic fibroblast) cells in response to UV-C irradiation. | 3 | 1.68E-03 | *RARA,EIF2C2,PARN* |
| WATANABE_COLON_CANCER_MSI_VS_MSS_DN | Down-regulated genes discriminating between MSI (microsatellite instability) and MSS (microsatellite stability) colon cancers. | 2 | 2.15E-03 | *EIF2C2,DACH1* |
| PID_SMAD2_3NUCLEARPATHWAY | Regulation of nuclear SMAD2/3 signaling | 2 | 2.20E-03 | *MYC,SNIP1* |
| ROSS_ACUTE_MYELOID_LEUKEMIA_CBF | Top 100 probe sets for core-binding factor (CBF) acute myeloid leukemia (AML): contains CBFB MYH11 [GeneID=865;4629] or AML1 ETO [GeneID=861;862] fusions. | 2 | 2.20E-03 | *BLVRA,EGFL7* |
| GGGAGGRR_V$MAZ_Q6 | Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif GGGAGGRR which matches annotation for MAZ: MYC-associated zinc finger protein (purine-binding transcription factor) | 7 | 2.33E-03 | *DACH1,MYC,RARA,BCL11B,MAFA,FMO3,TMCC3* |

## Supplementary Table 17. Geneset Enrichment Analyses of Squamous Cell Carcinoma versus Adenocarcinoma of the Cervix

### Genesets enriched in Squamous cell carcinoma

| GENESET | FDR q-val |
|---|---|
| JAEGER_METASTASIS_DN | 0 |
| ONDER_CDH1_TARGETS_2_DN | 0 |
| RICKMAN_TUMOR_DIFFERENTIATED_WELL_VS_POORLY_DN | 0 |
| CHR1Q21 | 0 |
| HUPER_BREAST_BASAL_VS_LUMINAL_UP | 0 |
| SMID_BREAST_CANCER_BASAL_UP | 0.00534398 |
| ZWANG_TRANSIENTLY_UP_BY_1ST_EGF_PULSE_ONLY | 0.007418505 |
| SMID_BREAST_CANCER_LUMINAL_B_DN | 0.03429207 |
| SENGUPTA_NASOPHARYNGEAL_CARCINOMA_WITH_LMP1_DN | 0.13030669 |

### Genesets enriched in Adenocarcinoma

| GENESET | FDR q-val |
|---|---|
| SMID_BREAST_CANCER_BASAL_DN | 0.064654365 |
| SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN | 0.20927976 |

## Supplementary Table 18. Genome rearrangements identified by whole genome sequencing

| Subject | Inter-chromosomal | Long-range (1Mb+) | Mid-range (10Kb–1Mb) | Local (<10Kb) | Total | Histology | Grade |
|---|---|---|---|---|---|---|---|
| SGCX-NOR-031 | 1 | 0 | 0 | 0 | 1 | Squamous cell carcinoma | 2 |
| SGCX-NOR-002 | 0 | 1 | 1 | 1 | 3 | Squamous cell carcinoma | 2 |
| SGCX-NOR-088 | 0 | 1 | 2 | 0 | 3 | Adenosquamous carcinoma | 3 |
| SGCX-NOR-093 | 1 | 1 | 1 | 2 | 5 | Squamous cell carcinoma | 2 |
| SGCX-NOR-104 | 2 | 3 | 1 | 0 | 6 | Adenocarcinoma | 1 |
| SGCX-NOR-025 | 3 | 1 | 4 | 1 | 9 | Clear cell carcinoma | 3 |
| SGCX-NOR-075 | 4 | 1 | 1 | 5 | 11 | Adenocarcinoma | 3 |
| SGCX-NOR-062 | 3 | 1 | 5 | 3 | 12 | Squamous cell carcinoma | 3 |
| SGCX-NOR-059 | 2 | 0 | 12 | 13 | 27 | Squamous cell carcinoma | 3 |
| SGCX-NOR-047 | 2 | 9 | 7 | 21 | 39 | Adenocarcinoma | 2 |
| SGCX-NOR-075 | 6 | 10 | 12 | 22 | 50 | Adenocarcinoma | 3 |
| SGCX-NOR-026 | 15 | 9 | 58 | 11 | 93 | Adenosquamous carcinoma | NA |
| SGCX-NOR-056 | 10 | 13 | 67 | 17 | 107 | Squamous cell carcinoma | 2 |
| SGCX-NOR-030 | 22 | 11 | 79 | 49 | 161 | Squamous cell carcinoma | 2 |

## Supplementary Table 19. Exon deletions and protein fusions observed in whole genome sequencing of 14 cervical carcinomas

EXON DELETIONS

| | chr1 | str1 | pos1 | chr2 | str2 | pos2 | #T | #N | class | fusion details | exons deleted |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SGCX-NOR-030 | chr20 | (+) | 15,206,279 | chr20 | (–) | 15435424 | 21 | 0 | deletion | *Deletion of 2 exons: in frame (MACROD2)* | 6-7 |
| SGCX-NOR-030 | chr20 | (+) | 14,630,754 | chr20 | (–) | 15063727 | 12 | 0 | deletion | *Deletion of 1 exon: in frame (MACROD2)* | 5-6 |
| SGCX-NOR-059 | chr11 | (+) | 70,640,928 | chr11 | (–) | 70645978 | 12 | 0 | deletion | *Deletion of 1 exon: in frame (SHANK2)* | 11 |
| SGCX-NOR-059 | chr2 | (+) | 20,196,627 | chr2 | (–) | 20200468 | 12 | 0 | deletion | *Deletion of 2 exons: in frame (MATN3)* | 5-6 |
| SGCX-NOR-026 | chr3 | (+) | 175,236,207 | chr3 | (–) | 175365542 | 8 | 0 | deletion | *Deletion of 2 exons: in frame (NAALADL2)* | 10-11 |
| SGCX-NOR-030 | chr2 | (+) | 133,613,206 | chr2 | (–) | 133618696 | 8 | 0 | deletion | *Deletion of 1 exon: in frame (NCKAP5)* | 11 |
| SGCX-NOR-056 | chr2 | (+) | 141,143,715 | chr2 | (–) | 141428935 | 7 | 0 | deletion | *Deletion of 25 exons: in frame (LRP1B)* | 42-66 |
| SGCX-NOR-030 | chr18 | (+) | 6,079,250 | chr18 | (–) | 6151068 | 7 | 0 | deletion | *Deletion of 3 exons: in frame (L3MBTL4)* | 14-16 |
| SGCX-NOR-030 | chr1 | (+) | 72,063,833 | chr1 | (–) | 72565489 | 6 | 0 | deletion | *Deletion of 4 exons: in frame (NEGR1)* | 2-5 |
| SGCX-NOR-030 | chrX | (+) | 96,185,328 | chrX | (–) | 96392965 | 6 | 0 | deletion | *Deletion of 12 exons: in frame (DIAPH2)* | 10-21 |
| SGCX-NOR-056 | chr3 | (+) | 56,880,303 | chr3 | (–) | 56967942 | 6 | 0 | deletion | *Deletion of 1 exon: in frame (ARHGEF3)* | 4 |
| SGCX-NOR-030 | chr2 | (+) | 141,982,136 | chr2 | (–) | 142058307 | 6 | 0 | deletion | *Deletion of 3 exons: in frame (LRP1B)* | 4-6 |
| SGCX-NOR-056 | chr9 | (+) | 87,351,402 | chr9 | (–) | 87425793 | 5 | 0 | deletion | *Deletion of 3 exons: in frame (NTRK2)* | 12-14 |
| SGCX-NOR-056 | chr7 | (+) | 157,421,645 | chr7 | (–) | 157462904 | 5 | 0 | deletion | *Deletion of 1 exon: in frame (PTPRN2)* | 13 |
| SGCX-NOR-056 | chr5 | (+) | 37,027,599 | chr5 | (–) | 37055840 | 5 | 0 | deletion | *Deletion of 10 exons: in frame (NIPBL)* | 33-42 |
| SGCX-NOR-056 | chr1 | (+) | 50,994,959 | chr1 | (–) | 51419719 | 4 | 0 | deletion | *Deletion of 14 exons: in frame (FAF1)* | 2-15 |
| SGCX-NOR-056 | chr5 | (+) | 58,852,572 | chr5 | (–) | 59255366 | 4 | 0 | deletion | *Deletion of 2 exons: in frame (PDE4D)* | 2-3 |
| SGCX-NOR-030 | chr2 | (+) | 133,631,068 | chr2 | (–) | 133674626 | 4 | 0 | deletion | *Deletion of 1 exon: in frame (NCKAP5)* | 9 |
| SGCX-NOR-059 | chr8 | (+) | [~ 98,788,343] | chr8 | (–) | [~ 98,828,281] | 4 | 0 | deletion | *Deletion of 2 exons: in frame (LAPTM4B)* | 2-3 |
| SGCX-NOR-026 | chr1 | (+) | [~ 193,105,257] | chr1 | (–) | [~ 193,138,790] | 4 | 0 | deletion | *Deletion of 5 exons: in frame (CDC73)* | 6-10 |
| SGCX-NOR-026 | chr14 | (+) | [~ 81,005,174] | chr14 | (–) | [~ 81,358,198] | 4 | 0 | deletion | *Deletion of 13 exons: in frame (C14orf145)* | 8-20 |

FUSIONS

| | chr1 | str1 | pos1 | chr2 | str2 | pos2 | #T | #N | class | *fusion details* |
|---|---|---|---|---|---|---|---|---|---|---|
| SGCX-NOR-030 | chr12 | (+) | 26,109,552 | chr12 | (+) | 26493110 | 19 | 0 | inversion | *Protein fusion: mid-exon (RASSF8-ITPR2)* |
| SGCX-NOR-026 | chr17 | (–) | 18,023,710 | chr17 | (+) | 18105432 | 7 | 0 | tandem duplication | *Protein fusion: mid-exon (ALKBH5-MYO15A)* |
| SGCX-NOR-104 | chr2 | (–) | 111,596,674 | chr2 | (+) | 143925356 | 7 | 0 | long range | *Protein fusion: in frame (ARHGAP15-ACOXL)* |
| SGCX-NOR-030 | chr3 | (+) | 119,048,894 | chr14 | (+) | 37591833 | 6 | 0 | inter chromosomal | *Protein fusion: in frame (ARHGAP31-SLC25A21)* |
| SGCX-NOR-026 | chr14 | (+) | 67,549,889 | chr14 | (–) | 67750786 | 6 | 0 | deletion | *Protein fusion: in frame (GPHN-MPP5)* |
| SGCX-NOR-026 | chr3 | (–) | 49,140,008 | chr3 | (+) | 49315808 | 5 | 0 | tandem duplication | *Protein fusion: mid-exon (QARS-USP4)* |
| SGCX-NOR-026 | chr8 | (–) | 121,224,039 | chr8 | (+) | 121480188 | 4 | 0 | tandem duplication | *Protein fusion: in frame (MTBP-COL14A1)* |
| SGCX-NOR-047 | chr22 | (–) | 26,383,489 | chr22 | (+) | 29649503 | 5 | 0 | long range | *Protein fusion: in frame (EMID1-MYO18B)* |
| SGCX-NOR-075 | chr11 | (+) | 130,012,116 | chr11 | (–) | 130050439 | 5 | 0 | deletion | *Protein fusion: in frame (APLP2-ST14)* |
| SGCX-NOR-075 | chr10 | (–) | 61,084,162 | chr10 | (–) | 64235755 | 4 | 0 | long range | *Protein fusion: in frame (FAM13C-ZNF365)* |

**chr1 str1 pos1 site1** = genomic position of one side of the rearrangement
**chr2 str2 pos2 site2** = genomic position of the other side of the rearrangement

**#T** = number of supporting readpairs in the tumor
**#N** = number of supporting readpairs in the normal

# Supplementary Table 20. Fusion events detected by RNA sequencing of 79 cervical carcinomas

| ID | Gene 1 | Gene 2 | Chimeric Pairs | Chr1 | Pos1 | Or1 | Chr2 | Pos2 | Or2 | DFRC | Interchromosomal | Sample had WGS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SGCX-NOR-028 | *RDH13* | *TIPARP* | 2 | 19 | 55568021 | R | 3 | 156413654 | F | 8 | Y | **N** |
| SGCX-NOR-030 | *IL20RB* | *RASSF8* | 32 | 3 | 136714398 | F | 12 | 26147970 | F | 8 | Y | Y* |
| SGCX-NOR-030 | *IL20RB* | *RASSF8* | 32 | 3 | 136714398 | F | 12 | 26173078 | F | 4 | Y | Y* |
| SGCX-NOR-030 | *PDE4A* | *MIPOL1* | 11 | 19 | 10574651 | F | 14 | 38016110 | F | 7 | Y | Y* |
| SGCX-NOR-036 | *PRSS3* | *PRSS1* | 2 | 9 | 33798080 | F | 7 | 142460282 | F | 4 | Y | N |
| SGCX-NOR-043 | *P4HB* | *OTUD7B* | 11 | 17 | 79813018 | R | 1 | 149949511 | R | 4 | Y | N |
| SGCX-NOR-048 | *ZNF532* | *RPA3* | 10 | 18 | 56532811 | F | 7 | 7677603 | R | 4 | Y | N |
| SGCX-NOR-080 | *CTSH* | *DEFB118* | 12 | 15 | 79237233 | R | 20 | 29960660 | F | 9 | Y | N |
| SGCX-NOR-080 | *HIBCH* | *CUX2* | 2 | 2 | 191152312 | R | 12 | 111772245 | F | 5 | Y | N |
| SGCX-NOR-088 | *DOCK1* | *SPECC1* | 8 | 10 | 128798571 | F | 17 | 20013740 | F | 15 | Y | **Y** |
| SGCX-NOR-043 | *ERBB2* | *C17orf37* | 744 | 17 | 37864787 | F | 17 | 37885858 | R | 5 | N | N |
| SGCX-NOR-043 | *ERBB2* | *IKZF3* | 305 | 17 | 37872192 | F | 17 | 37944627 | R | 10 | N | N |
| SGCX-NOR-059 | *AKR1B15* | *AKR1B10* | 155 | 7 | 134234001 | F | 7 | 134215395 | F | 0 | N | Y |
| SGCX-NOR-030 | *MYO15B* | *NUP85* | 70 | 17 | 73588382 | F | 17 | 73221198 | F | 12 | N | Y* |
| SGCX-NOR-083 | *SAA2* | *SAA1* | 46 | 11 | 18267457 | R | 11 | 18291264 | F | 2 | N | N |
| SGCX-NOR-104 | *PSMA3* | *ACTR10* | 41 | 14 | 58718960 | F | 14 | 58701088 | F | 8 | N | Y |
| SGCX-NOR-078 | *IMPDH2* | *WDR6* | 27 | 3 | 49065864 | R | 3 | 49049068 | F | 5 | N | N |
| SGCX-NOR-065 | *TFG* | *GPR128* | 20 | 3 | 100438902 | F | 3 | 100348442 | F | 6 | N | N |
| SGCX-NOR-005 | *STAP2* | *KDM4B* | 19 | 19 | 4338649 | R | 19 | 5150369 | F | 4 | N | N |
| SGCX-NOR-065 | *FAM169A* | *ANKRD31* | 19 | 5 | 74134790 | R | 5 | 74380199 | R | 10 | N | N |
| SGCX-NOR-086 | *ZNF28* | *ZNF468* | 15 | 19 | 53303301 | R | 19 | 53344252 | R | 1 | N | N |
| SGCX-NOR-048 | *KITLG* | *PTPRR* | 12 | 12 | 88974041 | R | 12 | 71056385 | R | 6 | N | N |
| SGCX-NOR-103 | *RACGAP1* | *CERS5* | 11 | 12 | 50419181 | R | 12 | 50528485 | R | 5 | N | N |
| SGCX-NOR-002 | *LAP3* | *CLRN2* | 8 | 4 | 17598757 | F | 4 | 17528440 | F | 8 | N | Y |
| SGCX-NOR-002 | *ZNF410* | *PTGR2* | 8 | 14 | 74363237 | F | 14 | 74340726 | F | 5 | N | Y |
| SGCX-NOR-030 | *TFG* | *GPR128* | 8 | 3 | 100438902 | F | 3 | 100348442 | F | 5 | N | Y |
| SGCX-NOR-085 | *ZNF33B* | *ZNF33A* | 8 | 10 | 43090043 | R | 10 | 38353016 | F | 3 | N | N |
| SGCX-NOR-080 | *FDPS* | *FMO3* | 7 | 1 | 155282186 | F | 1 | 171061794 | F | 5 | N | N |
| SGCX-NOR-006 | *PPP6R3* | *LRP5* | 6 | 11 | 68287119 | F | 11 | 68197043 | F | 7 | N | N |
| SGCX-NOR-056 | *STRN* | *HEATR5B* | 6 | 2 | 37193373 | R | 2 | 37195744 | R | 10 | N | Y |
| SGCX-NOR-090 | *ABI1* | *PDSS1* | 6 | 10 | 27149676 | R | 10 | 27031426 | F | 11 | N | N |
| SGCX-NOR-091 | *RREB1* | *LY86* | 6 | 6 | 7108293 | F | 6 | 6625159 | F | 5 | N | N |
| SGCX-NOR-006 | *AKR1C2* | *AKR1C1* | 5 | 10 | 5037511 | R | 10 | 5019892 | F | 0 | N | N |
| SGCX-NOR-021 | *NPEPPS* | *KIAA1267* | 5 | 17 | 45668247 | F | 17 | 44172067 | R | 5 | N | N |
| SGCX-NOR-026 | *STAT3* | *PTRF* | 5 | 17 | 40540297 | R | 17 | 40557406 | R | 3 | N | Y* |
| SGCX-NOR-028 | *ARL2* | *C2CD3* | 5 | 11 | 64786205 | F | 11 | 73825638 | R | 4 | N | N |
| SGCX-NOR-075 | *ZNF805* | *ZNF264* | 5 | 19 | 57755417 | F | 19 | 57722722 | F | 3 | N | Y |
| SGCX-NOR-083 | *DAZAP1* | *MUM1* | 5 | 19 | 1407801 | F | 19 | 1376518 | F | 5 | N | N |
| SGCX-NOR-103 | *RIPK1* | *SERPINB9* | 5 | 6 | 3064293 | F | 6 | 2900855 | R | 8 | N | N |
| SGCX-NOR-029 | *RAET1L* | *ULBP2* | 4 | 6 | 150342041 | R | 6 | 150269859 | F | 6 | N | N |
| SGCX-NOR-043 | *STARD3* | *PPP1R1B* | 4 | 17 | 37793484 | F | 17 | 37790136 | F | 7 | N | N |
| SGCX-NOR-065 | *BTN3A3* | *BTN3A1* | 4 | 6 | 26448676 | F | 6 | 26410122 | F | 1 | N | N |
| SGCX-NOR-079 | *C12orf62* | *SMARCD1* | 4 | 12 | 50506084 | F | 12 | 50488220 | F | 13 | N | N |
| SGCX-NOR-080 | *MFSD6* | *C2orf88* | 4 | 2 | 191280139 | F | 2 | 190944680 | F | 5 | N | N |
| SGCX-NOR-011 | *CRHR1* | *KIAA1267* | 3 | 17 | 43699407 | F | 17 | 44249598 | R | 3 | N | N |
| SGCX-NOR-029 | *AKR7L* | *AKR7A3* | 3 | 1 | 19596077 | R | 1 | 19610619 | R | 3 | N | N |
| SGCX-NOR-048 | *C5orf42* | *NUP155* | 3 | 5 | 37238959 | R | 5 | 37314430 | R | 6 | N | N |
| SGCX-NOR-052 | *SIPA1L1* | *RAD51B* | 3 | 14 | 71996087 | F | 14 | 69196525 | F | 4 | N | N |
| SGCX-NOR-085 | *BFSP1* | *PAK7* | 3 | 20 | 17489534 | R | 20 | 9525141 | R | 8 | N | N |
| SGCX-NOR-103 | *EP400NL* | *EP400* | 3 | 12 | 132593227 | F | 12 | 132472250 | F | 0 | N | N |
| SGCX-NOR-028 | *TYW1* | *CDK14* | 2 | 7 | 66532390 | F | 7 | 90836480 | F | 5 | N | N |
| SGCX-NOR-048 | *CRHR1* | *KIAA1267* | 2 | 17 | 43699407 | F | 17 | 44249598 | R | 4 | N | N |
| SGCX-NOR-065 | *CRHR1* | *KIAA1267* | 2 | 17 | 43699407 | F | 17 | 44249598 | R | 3 | N | N |
| SGCX-NOR-066 | *ZNF33B* | *ZNF33A* | 2 | 10 | 43090043 | R | 10 | 38353016 | F | 1 | N | N |
| SGCX-NOR-073 | *ZDHHC11B* | *ZDHHC11* | 2 | 5 | 733867 | R | 5 | 824208 | R | 3 | N | N |
| SGCX-NOR-076 | *CRHR1* | *KIAA1267* | 2 | 17 | 43699407 | F | 17 | 44249598 | R | 5 | N | N |
| SGCX-NOR-085 | *AKR7L* | *AKR7A3* | 2 | 1 | 19596077 | R | 1 | 19610619 | R | 1 | N | N |
| SGCX-NOR-096 | *AKR7L* | *AKR7A3* | 2 | 1 | 19596077 | R | 1 | 19610619 | R | 1 | N | N |
| SGCX-NOR-103 | *LGALS7B* | *CAPN12* | 2 | 19 | 39281530 | F | 19 | 39233168 | R | 4 | N | N |

*Chr1 Pos1 Or1: Chromosome, Position and Orientation of fusion gene 1*
*Chr2 Pos2 Or2: Chromosome, Position and Orientation of fusion gene 2*
*DFRC : Distinct Fusion Reads Consistent (distinct fusion reads in orientation consistent with their mate)*
*FRG1/FRG1B and HLA/HLA fusions were removed from this table as they likely represent alignment artifacts (See Supplementary Note 13B)*
** Denotes events supported by fusion calls in WGS*

## Supplementary Table 21. Mutations in the APOBEC Gene Family

| Patient | Histology | Tumor grade | Nonsilent mutation rate (/Mb) | Relative frequency of Tp*C mutations | Mutated APOBEC family gene | Mutation |
|---|---|---|---|---|---|---|
| SGCX-NOR-055 | Squamous cell carcinoma | 2 | 39.49 | 0.898 | *APOBEC3B*<br>*APOEBEC3F* | S93F<br>S92F |
| SGCX-MEX-001 | Squamous cell carcinoma | 3 | 31.42 | 0.863 | *APOBEC3G* | L189M |
| SGCX-MEX-006 | Squamous cell carcinoma | 2 | 7.11 | 0.752 | *APOBEC3G*<br>*APOBEC3G* | D317N<br>E323Q |
| SGCX-NOR-048 | Squamous cell carcinoma | 2 | 6.87 | 0.712 | *APOBEC3G* | E217G |

## Supplementary Table 22b. All Mutations in Fanconi Anemia Genes

| Gene | Somatic Non-silent Mutations (N) | Germline Non-silent Mutations (N) | Number of Individuals |
|---|---|---|---|
| *BRCA2* | D1769G(1), E2875K(1), S2697R(1) | A2951T(3), S1760A(1), I2490T(6), R2034C(2), T1915M(3), D596H(2), V2466A(115), N991D(6), I1929V(1), N289H(6), I3412V(3), K3326*(3), D1420Y(1), N372H(54) | 115 |
| *BRIP1* | D149H(1) | K297R(2), S919P(94), V193I(4), R798*(1) | 95 |
| *FANCA* | E629K(1), S175_splice(1), A444V(1) | W745L(2), T266A(77), S1088F(15), A181V(1), P201S(1), M717I(9), G501S(77), A412V(15), S858R(2), G809D(66) | 102 |
| *FANCB* |  | G335E(15), I330T(2) | 17 |
| *FANCC* |  | A505T(1), D195V(1) | 2 |
| *FANCD2* | I796_splice(1) | G901V(9) | 10 |
| *FANCE* | P211S(1) | S204L(3), A502T(18) | 19 |
| *FANCF* |  | P320L(4), V295I(1), A186V(1) | 6 |
| *FANCG* |  | R513Q(4) | 4 |
| *FANCI* | L312V(1), R514K(1), G896E(1) | P55L(9), A86V(76), I132V(1), G422R(1), L605F(2), P471R(1), M525V(2), C742S(78) | 87 |
| *FANCL* |  | E147K(1), L38F(1), M247V(1) | 3 |
| *FANCM* | R756H(1), D53Y(1) | I208M(3), I349T(1), K953N(1), I1460V(23), P90L(1), S175F(7), L57F(3), R1931*(1), P1812A(23), N655S(4), T1600I(5), N1253S(1), V878L(24), N1876S(1), Q1701*(1), I259V(1), T77A(1), I1742V(1), Y413H(1) | 42 |
| *PALB2* | D772N(1) | 19G(1), P864S(1), G998E(1), Q559R(14), V932M( | 28 |
| *RAD51C* |  | G264S(3), T287A(1) | 4 |

## Supplementary Table 22b. Fanconi Anemia Genes with Both Somatic and Germline Mutations

| Gene | Individual | Somatic Non-silent Mutations | Germline Non-silent Mutations |
|---|---|---|---|
| *BRCA2* | SGCX-NOR-048 | S2697R | V2466A |
| *BRCA2* | SGCX-NOR-026 | D1769G | N372H, V2466A |
| *BRCA2* | SGCX-NOR-011 | E2875K | N372H, V2466A |
| *FANCA* | SGCX-NOR-045 | A444V | G809D, T266A |
| *FANCA* | SGCX-NOR-055 | E629K, S175_splice | G809D, G501S, T266A |
| *FANCI* | SGCX-NOR-094 | G896E | A86V, C742S |
| *FANCI* | SGCX-NOR-043 | L312V | A86V, C742S |