Supplementary Materials and Methods:

Supplementary Methods

Cervical cancer and matched normal samples

Chinese cervical carcinoma patients were recruited at Southwest Hospital (SWH). Patient informed consent was provided at the hospital under the supervision of the hospital's institutional review board. Peripheral blood samples and surgically resected tumor and normal tissues were collected from each patient. The adjacent normal and tumor tissues was separated during surgery. And samples were confirmed by two pathologists using H&E staining. The tissue samples were immediately immersed in RNAlater solution (Life Technologies, CA, USA) and stored at 4°C overnight to ensure that the solution penetrated the tissue before being stored at -80°C. Blood samples were stored at -20°C. A total of 168 sample pairs were collected at the hospital. Cases were staged according to the FIGO staging system in the contributing hospital. Clinical data, including age, pathological type, stage, grade, and other information, were collected. An independent pathological review of the tumor and normal tissues was performed in another hospital by two independent pathologists to confirm the diagnosis and to analyze the tumor cell content. Tumor tissues with tumor cell content less than 40% and normal tissues containing tumor cells were excluded from further analysis. Thirty-nine sample pairs were excluded in this step. Germline DNA was obtained primarily from peripheral blood and/or from matched normal tissues when no peripheral blood sample was available. This project and protocols involving human and animal tissues

were approved by the Research Ethics Committee of Southwestern Hospital (No. 2014-016) and conducted according to the Declaration of Helsinki.

Cases with qualified DNA and RNA samples were further included in the sequencing analysis (four sample pairs that failed the molecular quality check were removed). Additionally, genotyping of 21 SNP loci was performed to ensure the correct pairing of normal and tumor samples, and three sample pairs were removed because of discordant genotypes. A total of 112 sample pairs were used for further analysis after ten sample pairs were removed because of poor data quality.

Normal Cervical tissue cells and CIN specimens

In this study, the tissue cells of normal or CIN were collected from female patients attending a sexually transmitted disease clinic in Shanghai Skin Disease Hospital. The cervical tissue cells were sampled using cervical brushes for each enrolled patient. The The HPV-DNA testing were analyzed by HC2 method (Qiagen, Hilden, Germany) for these samples. And the HPV types were tested by Tellgen 2009-HPV DNA 26 typing kit (Tellgen Company, Shanghai, CN) for samples with HPV-positive. In addition, these samples were used to observe the morphological characteristics based on the liquid-based cytology (Becton Dickinson Company, New Jersey, USA). In Shanghai Skin Disease hospital, cervical liquid-based cytological test was used as a routine screening test against those patients who infected with HPV. Hence, those patients with HPV infection or with cytological abnormality were recommended to accept colposcopy examination. And then the colposcopy findings were used to determine if a

biopsy is necessary. The diagnoses of normal, CIN1, CIN2 or CIN3 were made and reviewed by two pathologists independently.

For individuals with normal cervical cytology, colposcopy and biopsy are not recommended[1-3]. However, colposcopy and biopsy were also performed in this study in the following situations: 1) Patients were very anxious about their condition (after high-risk sexual activity) and insisted on accepting further examinations, including biopsy; and 2) Patients agreed to be involved in this study after informed consent was signed. Hence, all patients who were recruited into this study have accepted the examinations of colposcopy and biopsy. The samples of normal or CINs were confirmed by histopathology, which were then used for analysis of HPV integration. The study was approved by the Ethics Committee of the Shanghai Skin Disease Hospital (No. SKIN2015-010).

DNA extraction and whole-genome or whole-exome sequencing

Genomic DNA was extracted from the blood and tissue samples using TIANamp Blood DNA Kits and TIANamp Genomic DNA Kits (Tiangen Biotech, Beijing, China), respectively, according to the manufacturer's protocols. The amount and integrity of the DNA were assayed using a Qubit® 2.0 Fluorometer (Life technologies, CA, USA) and gel electrophoresis.

WGS was performed on paired cancer and normal tissue samples using the complete genomics (BGI, Shenzhen, China) [4] and HiSeq X-Ten (Illumina, CA, USA) sequencing platforms. For the CG platform, genomic DNA was fragmented, amplified and used to

generate DNA nanoballs, which were then placed on different sticky spots on a silicon chip to form the so-called DNA nanoball arrays. Sequencing was then performed using combinatorial probe-anchor ligation technology to generate 30-bp paired-end reads. For the HiSeq X-Ten platform, the DNA was fragmented and selected (approximately 500 bp), and then three enzymatic steps (end repairing, the addition of an "A" base, and adapter ligation) were performed to generate the library.

WES was performed on HiSeq 2000 (Illumina, CA, USA) and Ion proton (Life Technologies, CA, USA) platforms. For the HiSeq 2000 platform, genomic DNA from matched tissues and control samples were fragmented using a bioruptor UCD-200 (Diagenode, NJ, USA) with a peak of 250 bp. Three enzymatic steps (end repairing, the addition of an "A" base, and adapter ligation) were performed to generate the library following the instructions from Illumina. Whole-exome enrichment was performed using the SeqCap EZ Human Exome Library v3.0 (Nimblegen, WI, USA). High-throughput sequencing was performed on an Illumina HiSeq 2000/2500 platform to generate 101 bp paired-end reads. For the Ion proton platform, cancer-normal paired DNA samples were enriched for exome sequencing using an AmpliSeq exome kit (Life Technologies, CA, USA), and libraries were constructed following a standard Ion Proton library construction protocol.

Collectively, the average amount of generated data per sample in giga base pairs (Gbp) and the paired-end libraries for each sequencing platform are CG sequencing (320Gbp per sample and 30bp read length), Illumina whole-genome sequencing (~172Gbp per sample and 150bp read length), Illumina whole-exome sequencing (~19Gbp data per

sample and 101bp or 150bp read length), and Ion Proton sequencing (13Gbp per sample and unequal read length ranging from 86-222bp), respectively.

RNA extraction and sequencing

RNA from cervical carcinoma and matched normal tissue samples was extracted using Trizol reagent according to the manufacturer's instructions (Life Technologies, CA, USA). Quality control for the RNA samples was performed using a 2100 Bioanalyzer Instrument (Agilent Technologies, CA, USA) to assay the amount and integrity of the RNA. Then, RNA sequencing libraries were constructed using TruSeq RNA Sample Prep Kits, v2 (Illumina, CA, USA) and sequenced using a HiSeq 2000 sequencer. Approximately 8Gbp were generated for each sample by RNAseq.

HPV typing and HPV probe capture sequencing by HPCS

HPV status was determined using mass spectrometry (MS) to analyze 14 high-risk HPV types (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, and 68) and two low-risk HPV types (HPV6 and 11) according to previous descriptions [5].

HPCS consists of three parts: sequencing library construction, HPV sequence enrichment, and next-generation sequencing. Sequencing library construction was performed exactly as for HiSeq 2000 WES, which was carried out by following the instructions from Illumina. DNA was sheared to approximately 250 bp using a bioruptor UCD-200 (Diagenode, NJ, USA) and was then end-blunted, A-tailed, and adapter-ligated. DNA was purified after each step with Ampure beads (Beckman Coulter, CA, USA). The concentrations of the libraries were assessed with a 2100

Bioanalyzer Instrument (Agilent Technologies, CA, USA). For the CC samples, no additional libraries were constructed; the same library constructed for whole exome capture was used for HPV sequence capture.

An HPV probe set was designed to cover whole genomes of 40 HPV types (HPV reference genomes were obtained from NCBI: ftp://ftp.ncbi.nih.gov/genomes/) and synthesized by MyGenostics (MyGenostics, Beijing, China), and the enrichment of HPV sequences was performed using a GenCap enrichment kit according to the manufacturer's instructions. Three micrograms of library were hybridized with HPV probes at 65 °C for 24 hours and then washed to remove the un-captured fragments. The enriched sequences were purified and amplified by 16 PCR cycles.

The enriched HPV sequence from the human genomic DNA library was then sequenced using a HiSeq 2000 to yield approximately 1 million (M) paired-end reads (100 bp each). The HPV strains that were captured in this analysis included all 16 HPV strains that were detected using MS assays and 24 additional strains (HPV10, 26, 30, 32, 40, 42, 43, 44, 53, 61, 62, 67, 70, 72, 73, 74, 81, 83, 84, 86, 87, 89, 90, and 91).

Next generation sequencing data processing

Reads generated on the CG platform were processed and assembled using the complete data management solution of Complete Genomics.

Reads generated on the HiSeq 2000 and HiSeq X Ten platforms were processed using the same protocol. First, the raw sequencing reads were filtered as follows to obtain clean data: 1) reads polluted by adapter sequences were removed, as were their mate

pair reads; 2) reads with more than 50% of low-quality bases (Q<5) and their mate pair reads were removed; and 3) reads with a high N rate of more than 0.1 and their mate pair reads were removed. The clean reads were then aligned to hg19 (10 CCs and 25 CINs sequenced whole exome on Hiseq X Ten platform were aligned to hg38) using BWA (Burrows-Wheeler Aligner) [6]. Reads with multiple mapping loci in the genome and reads with more than 3 mismatches, with more than 1 gap, or with a gap more than 20 bases long were removed. The duplicated reads derived from PCR amplification were marked using Picard tools (http://broadinstitute.github.io/picard/). Local realignments and base quality recalibrations were performed using GATK [7].

The bam files generated using the Ion Proton sequencing platform were filtered as follows to obtain high-quality data: 1) reads less than 30 bp or more than 280 bp were removed; 2) reads with a start base not in any primer of the AmpliSeq exome kit primer set were removed; and 3) the first few bases of the reads that overlapped with the primer they were derived from were removed.

The final bam files that were generated were then used for further analysis. The depth of coverage of the target region and the fractions of regions with at least 4-fold, 10-fold, or 20-fold coverage were analyzed. The genotypes of the 21 common SNP sites were checked to ensure that the tissue and matched normal samples were derived from the same patient.

Mutation calling and gene mutational significance analysis

Somatic SNV were called using the Strelka [8] (for 10 CCs and 25 CINs sequenced on hiseq X Ten platform) or MuTect algorithm [9], and further filtrations were performed to obtain a high-confidence somatic SNV. The filter criteria for the SNVs that were called from the WES data generated on the HiSeq 2000 platform were as follows: 1) the reads according to the altered allele in the tumor were no less than five; 2) the fraction of reads according to the altered allele in the tumor was no less than 5%; and 3) the fraction of the reads according to the altered allele in the matched control was less than 3%. SNVs called from the WGS data generated on the HiSeq X Ten platform were filtered using criteria 2 and 3. The filter criteria for SNVs called from the WES data that were generated on the Ion Proton platform were as follows: 1) the depth of coverage was no less than 20 in the tumor samples and no less than 14 in the control samples; 2) the fraction of the reads supporting the mutation was no more than 2% in the control samples; 3) the fraction of the reads supporting the mutation in the tumor samples was five-fold or more than that in the control samples; 4) the reads supporting the alteration were either no less than 13 and present as a fraction no less than 5% or were no less than 5 and present as a fraction no less than 10%; 5) the mutation did not reside in a highly repeated region or a region with low sequence complexity; and 6) the mutation was supported by both directions of the relevant reads and did not show obvious strand bias (when the supporting reads for a minor allele were only 1, the fraction was no less than 20%, and the fraction was no less than 10% when the supporting reads were more than 1).

Somatic small insertions or deletions (indels) were detected using VarScan 2 tools [10] with the default parameters and then were subjected to 3 steps of filtering, as follows: 1) variants that were detected in the matched normal sample (with 3 or more reads supporting an indel in the normal sample at the same location as in the tumor sample or in the 40 bp flanking region) were removed; 2) variants residing near the polyN region were removed; and 3) variants residing near a region with low complexity or short tandem repeat regions were removed. The identified mutations were annotated for information related to the location, function, previous reports, and sequencing data supporting the status of the mutation using a local program. Somatic SNVs and indels for CG-sequenced samples were called and annotated using CG data management solutions. A MutsigCV tool [11] was used to find significantly mutated genes. Somatic SNVs and indels located in exons and spliced regions were input into MutsigCV and analyzed using the default parameters. Genes with a FDR no more than 0.1 were considered to be significantly mutated.

Copy number variation analysis

For the CG sequenced samples, CNAs were called using CG data management solutions, and for the other samples, CNAs were called using CLImAT software [12] with the input data processed using local scripts. For HiSeq X Ten-sequenced samples, the whole genome was divided into 400 bp bins that contained a SNP site and its flanking sequence. A total of 12,183,239 bins were generated for each sample using SNPs chosen from the dbSNP database (dbSNP 137) with a MAF of no less than 0.05 and no clustering in genomic regions. The read depth was then extracted for each bin from the

bam files, and the copy ratios were calculated by dividing the depth of each bin of cancer tissues by the depth of the same bin of matched controls. Then, we performed GC normalization before performing centralization using the copy ratio mode. The correlation between the copy ratio and the GC content was normalized, and the noise caused by GC bias was reduced. The B allele frequency was calculated and integrated with the final copy ratio as the input data for CLImAT. CNAs were called using this HMM-based method. CNAs from the WES samples were called using the same method except for the target region dividing process. Fragments of the target regions in the SeqCap EZ Human Exome Library v3.0 were used as the bins for the HiSeq 2000 cohort, while amplicons of the AmpliSeq exome kit were used as the bins for the Ion Proton cohort. Five HiSeq 2000 sequenced samples, including CC-H001, CC-H006 (the sample subjected to WES on both the HiSeq and the proton platforms), CC-H014, CC-H017 and CC-H021, were subjected to low-depth whole-genome sequencing. Correlation analyses between the CNAs generated from the WGS and WES data were performed to evaluate whether the somatic CNAs called using the WES data had a detecting power that was similar to that of the WGS data.

We analyzed multiploid genomes using a local script according to the results of CLImAT and the following criteria: 1) if >50% of the 'normal regions' showed LOH, we considered the sample to be multiploid; and 2) if >10 M deleted regions showed balanced allele-frequency, we considered the sample to be multiploid. GISTIC2.0 tools [13] were used to identify the frequently altered broad and focal CNAs ($q \leq 0.25$), and an in-house designed algorithm was used to find additional CNA peaks. Besides, an in-

house design algorithm was used to identify CNA peaks as a supplement, which contain four steps. 1, For each gene, the number of samples had an amplification or deletion on which was counted and then a p value based on one tail test of Poisson distribution was calculated both for amplification and deletion. 2, For each chromosome arm, we searched out local maximum value for amplification numbers and deletion numbers. A gene or several contiguous genes which has local maximum value was considered as a candidate peak if the p value of its amplification number/deletion number is no more than 0.01. 3, We discard a candidate peak if the following conditions were satisfied. (1), There is another candidate peak within ten megabase from this candidate peak and the amplification/deletion numbers of it was higher than that of this candidate peak (2), The average amplification/deletion numbers of genes between the two candidate peaks was higher than that of this candidate peak. 4, To get the peak regions for peaks passed the third step, we expand the coordinates on both sides until (1) p value of the gene on the border >0.01. (2) The difference of amplification/deletion numbers between the gene on the border and the peak was no more than five.

Genomic rearrangement and gene fusion analysis

We used the CREST algorithm [14] to detect structural variation in the WGS data describing the matched tumor and normal tissues. The CREST algorithm extracted soft-clipped reads to detect structural variation in sites at a base pair resolution. The software used bam files, which were generated by BWA alignment. We performed CREST using the default parameters. The pipeline included three steps: 1) we extracted soft-clipped breakpoints; 2) we removed germline variation by running the Perl script countdiff.pl;

and 3) we detected structural variations. The detection procedure was processed as follows: we assembled the soft-clipped reads, mapped the sequences, reassembled the other side of the putative breakpoint using BLAT (BLAST-like Sequence Alignment Tool), and, finally, if both sides showed high alignment similarity, they were reported as a putative structural variation event.

Chromosomes with inferred chromthripsis were detected based on the criteria proposed by Korbel and Campbell [15]. Briefly, evidence of chromothripsis was considered when breakpoints were highly clustered in a chromosome (exponential distribution, $p<0.001$), copy numbers oscillations$\geq$6, or when loss of heterogeneity, random DNA joins, or randomly ordered DNA fragments were observed.

To identify fusion genes, we used the TopHat-Fusion algorithm to detect the mRNA expression of fusion genes [16]. Putative fusion genes were filtered out by at least two span reads, at least two pair end reads, and no less than five total supporting reads.

HPV integration analysis

HPV sequences from human genomic DNA were captured and sequenced to an ultra-high coverage, yielding approximately one GB of data for each sample. On average, 94.8% of the raw data (total of 1.6 GB) after the removal of the low-quality data were used for HPV integration breakpoint analysis. Five steps were performed to analyze HPV integration, namely, reads identification, location and orientation definition, clustering and filtering of results, annotation, and integrant (integrated HPV fragments) prediction.

To identify reads that supported HPV integration, reads were aligned to the HPV genome and, subsequently, to the human genome using BWA. HPV reference genomes were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/). A single-end read of at least 50 bp mapped to the HPV genome was kept, as well as its matching pair reads, and these reads were then mapped to the human genome. Then, reads of at least 30 bp that matched and their matching pair reads were kept. Chimeric reads (when one part of a read mapped to the HPV genome and the other part mapped to the human genome) and discordant read-pairs (when one read mapped to the HPV genome and the matched read mapped to the human genome) were identified, and HPV integrations were analyzed using these reads. HPV integration has previously been shown to cause variation around the integration site [17]. Therefore, 10-bp mismatches or gaps were allowed for discordant read-pairs, and a 15-bp overlap was allowed for chimeric reads.

Exact locations in the human and HPV genomes were identified for chimeric read-supported integrations, and for discordant read-pairs, the positions nearest to the potential junction were reported. In addition, integration orientations were defined according to which end of a read mapped to which strand of the HPV and human genomes. The orientation of a chimeric read at the human end was considered (1) plus (+) if the chimeric read mapped first to the human genome or (2) minus (-) if the chimeric read did not map first to the human genome. In the viral end of the reads, the orientation was defined as plus (+) if a chimeric read was not first mapped to HPV and as minus (-) if a chimeric read was mapped first to HPV. The orientation of a discordant read-pair was defined as plus (+) when the read mapped to the forward strand of the

human or HPV genome and minus (-) when the read mapped to the reverse strand of the human or HPV genome.

Integration breakpoint clustering and filtering were performed as the third step. Chimeric reads that mapped as nearly identical (the predicted breakpoints were located within ±70 bp in the human genome and ±70 bp in the HPV genome and had the same orientation) were considered to be one integration event, with the leftmost integration site being the output. Then, integrations supported by discordant read-pairs were clustered to the adjacent integrations (located within 500 bp and having the same orientation) supported by chimeric reads, and locations of chimeric read-supported integrations were reported. Then, discordant read-pair-supported integrations that could not cluster to a chimeric read-supported integration were clustered with each other at locations within 600 bp in the human genome and 500 bp in the HPV genome and with the same orientation. The nearest positions to the predicated junctions were reported. After clustering, the results were filtered, with removal of integrations supported by only one read.

Annotation of high-confidence integrations was performed using ANNOVAR [18] and a local algorithm. The genomic location of the integration was annotated, including the genomic position and cytoband, and disrupted genes in both the human and HPV genomes (upstream and downstream genes were reported for the integrations located in intergenic regions).

To check the sizes of the viral integrants within the human genome, we identified the most possible integration partners for each integration junction. Two closet breakpoints

located in the same chromosome and same HPV strains genome, with complementary orientations, were identified as integration partners. Size of HPV integrants were calculated using these partners.

Validation for gene mutation and HPV integration

To validate the somatic SNVs identified in our cohort, mass spectrometry (MS)-based genotyping was used. Multiplex genotyping primers and extension probes were designed using Agena Bioscience's online design tools (https://mysequenom.com/Logon.aspx?ReturnUrl=%2fTools), and genotyping was performed on a MassArray® System (Agena Bioscience, Hamburg, Germany). Somatic status was evaluated by comparing the genotyping results from cancer DNA and their matched normal DNA samples. PCR amplification and Sanger sequencing were combined to validate SNVs and indels for which we could not find a proper MS primer pair.

HPV integration junctions and integrants were also validated using PCR and Sanger sequencing. For integration junctions, primers pairs were designed with one primer mapped to the human genome, the other primer mapped to the HPV genome, and with the junction located within the predicted amplicon. Successfully amplified samples were subjected to Sanger sequencing to confirm the integration breakpoints in both the HPV and the human genomes. For integrants, long PCR was used with both primers mapped to the human genome and with the full length of the integrated HPV fragment and the surrounding human sequence on both ends included within the predicated

amplicon. Sanger sequencing was performed on both ends of the amplicon to validate both junctions of the integrant.

RNAi-mediated gene silencing and cell viability analyses

Three siRNAs directed against 47 target genes were designed using the Whitehead Institute Web Server (http://jura.wi.mit.edu/bioc/siRNAext/) and then chemically synthesized (Shanghai GenePharma Co., Shanghai, China) to target different coding regions of each gene. The siRNA sequences are available upon request. In addition, siRNA-NC (5'-GAGUUAAAGUCAAAGUGACTT-3' and 5'-GUCACU UUGACUUUAACUCTT-3') was also synthesized. The siRNAs were transfected into cervical cancer cell lines and cell growth was monitored. For siRNA transfections, 3 × 103 cervical cancer cells (Hela229, C-33A, Caski and HeLa cells, purchased from China Infrastruture of Cell Line Resource) per well were seeded into 96-well plates. When the cells reached 30% to 50% confluence, they were transfected with synthetic siRNAs at a final concentration of 50 nM using the Lipofectamine 2000 Transfection Reagent (Invitrogen, CA, USA) according to the manufacturer's instructions.

The cells were cultured for 7 days and cell viability was measured using an ACEA RTCA kit (ACEA Biosciences, CA, USA). A microelectronic cell sensor system was used to confirm the number of living cells. Hela229, C-33A, Caski and HeLa cells (1 × 104) were seeded into each sensor-containing well (a 19.6-mm2 surface with 150 mL of medium) of the microtiter plates. The electronic sensors provided a continuous (every 6 h), quantitative measurement of the cell index (reflecting the surface area covered by the cells) in each well. Cell growth was measured every 6 h for 96 to 144 h,

and the cell indexes were recorded for each well at all time points to assess cell viability. All experiments were independently repeated at least three times.

References:

1. Cadman L. Colposcopy: A Practical Guide (2nd edn): British Medical Journal Publishing Group, 2013.
2. Huh WK, Ault KA, Chelmow D, Davey DD, Goulart RA, Garcia FA, Kinney WK, Massad LS, Mayeaux EJ, Saslow D, Schiffman M, Wentzensen N, Lawson HW, Einstein MH. Use of primary high-risk human papillomavirus testing for cervical cancer screening: interim clinical guidance. *Obstet Gynecol* 2015;125(2):330-7.
3. Bentley J, Bertrand M, Brydon L, Gagné H, Hauck B, Mayrand M-H, McFaul S, Power P, Schepansky A, Straszak-Suri M. Colposcopic management of abnormal cervical cytology and histology. *Journal of Obstetrics and Gynaecology Canada* 2012;34(12):1188-202.
4. Ciotlos S, Mao Q, Zhang RY, Li Z, Chin R, Gulbahce N, Liu SJ, Drmanac R, Peters BA. Whole genome sequence analysis of BT-474 using complete Genomics' standard and long fragment read technologies. *Gigascience* 2016;5:8.
5. Yi X, Li J, Yu S, Zhang A, Xu J, Yi J, Zou J, Nie X, Huang J, Wang J. A new PCR-based mass spectrometry system for high-risk HPV, part I: methods. *Am J Clin Pathol* 2011;136(6):913-9.
6. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997* 2013.
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20(9):1297-303.

8. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012;28(14):1811-7.

9. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31(3):213-9.

10. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22(3):568-76.

11. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau D-A, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-18.

12. Yu Z, Liu Y, Shen Y, Wang M, Li A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics* 2014;30(18):2576-83.

13. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;12(4):R41.

14. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 2011;8(8):652-4.

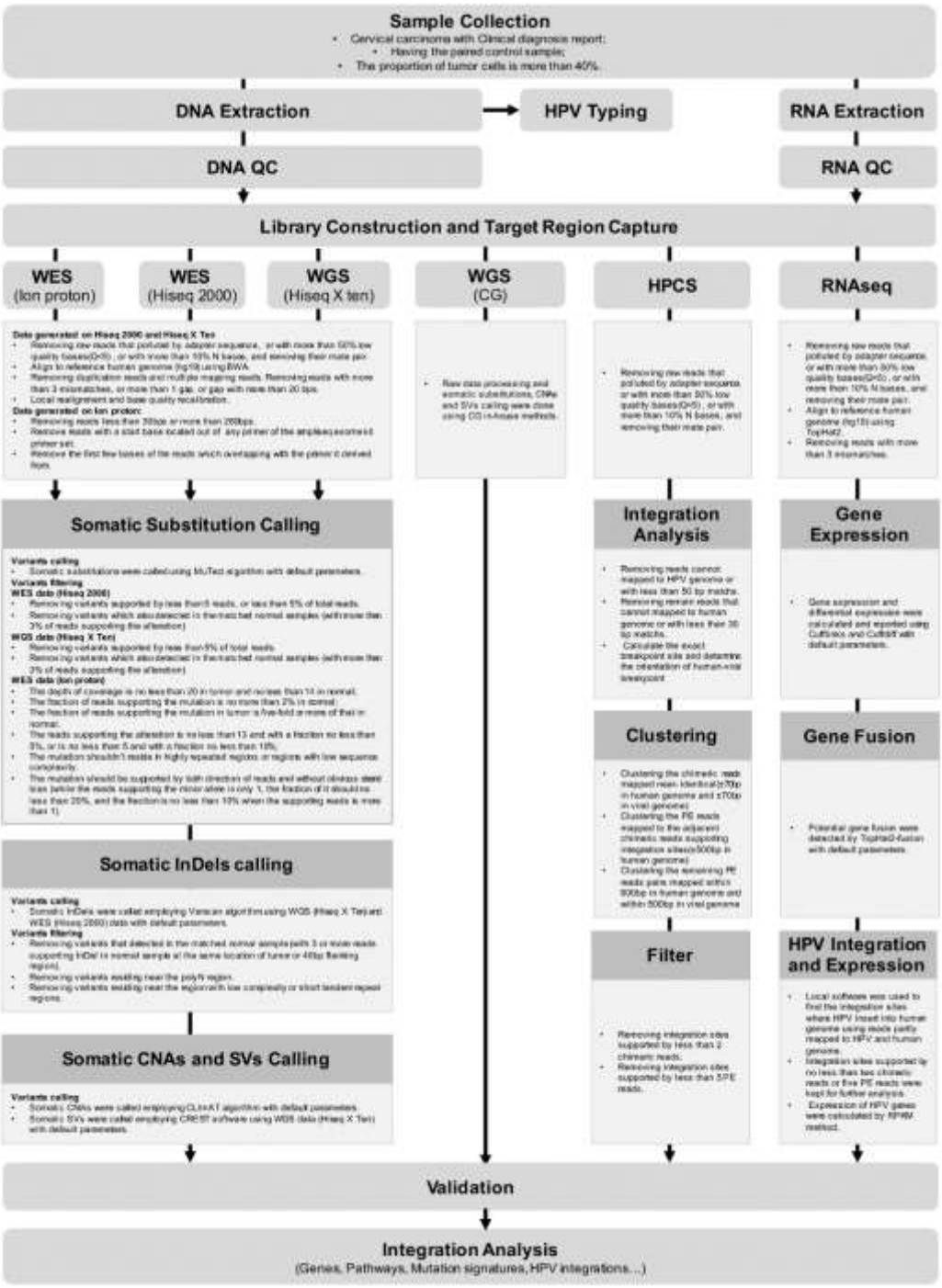
15. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell* 2013;152(6):1226-36.

16. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* 2011;12(8):R72.

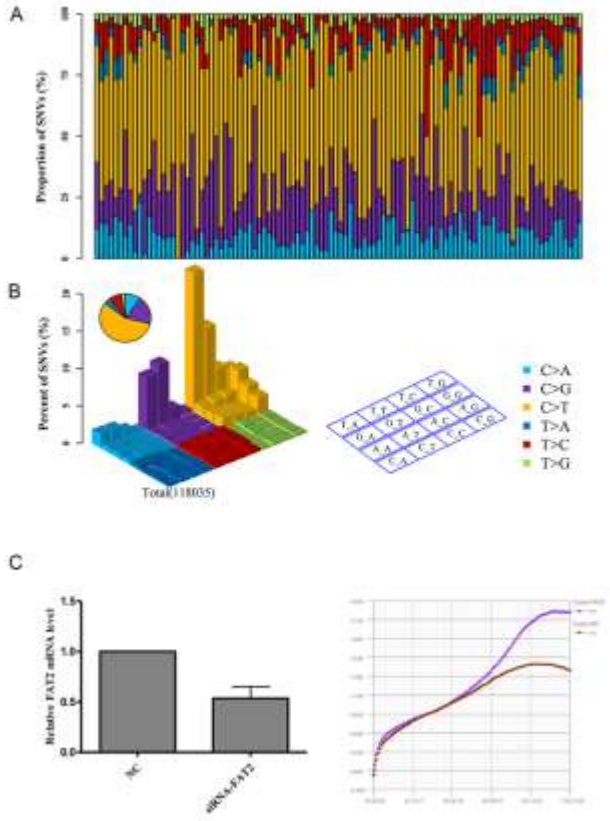
17. Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, Rocco JW, Teknos TN, Kumar B, Wangsa D, He D, Ried T, Symer DE, Gillison ML. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome research* 2014;24(2):185-99.

18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38(16):e164.

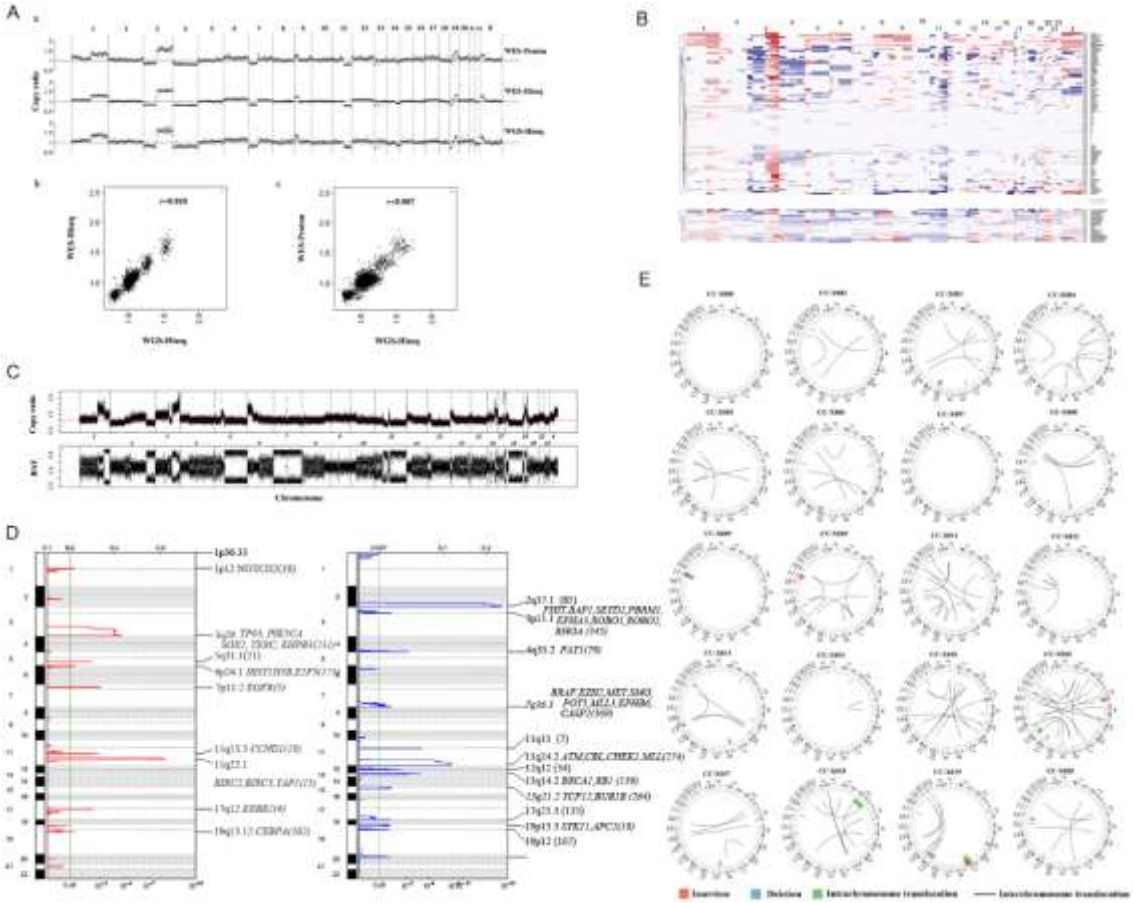
Supplementary figures



Supplementary Figure S1. Flow chart of sample collection, sequencing and data analysis.

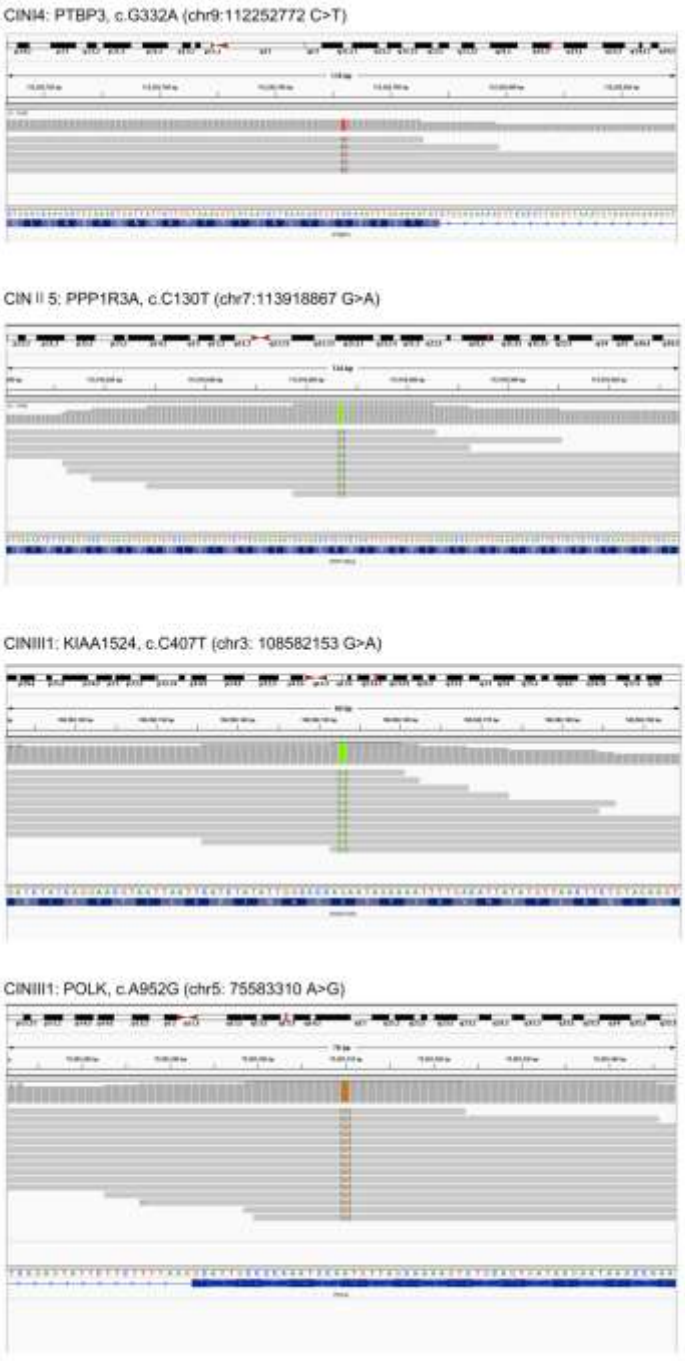


Supplementary Figure S2. Mutation signature and driver gene analysis. (A) The proportions of six mutation types in 102 of the samples are shown. Each vertical track represents one sample. The types of mutations are represented by different colors in the lower right. (B) The percent of 3-base mutation types in a total of 118,035 somatic SNVs in 102 CCs are shown. Each colored square shows the sequence context labeled using the legend on the right. The proportions of the six mutation types out of the total mutations are shown in the pie graph. (C) Relative FAT2 mRNA levels (a) and cell viability (b) in Caski cells treated with siRNA targeting the *FAT2* gene or a negative control siRNA (NC). Silencing of FAT2 using siRNA can significantly reduce FAT2 mRNA level ($p<0.05$, t-test), and promotes cell growth of Caski cells ($p<0.05$, t-test).

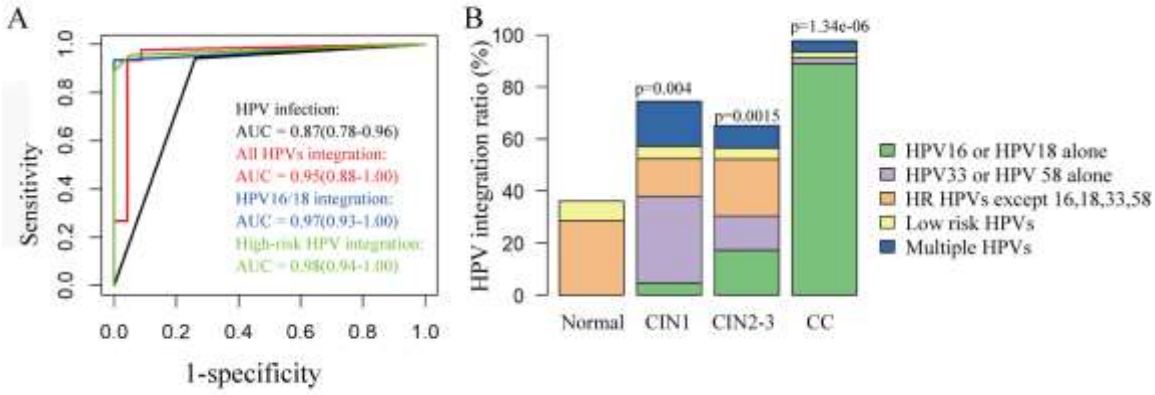


Supplementary Figure S3. Copy number alterations and structure variations analysis. (A) Correlation of CNA data in different platforms. The copy ratios of CC-H006/CC-P001 generated by the three different data sets (WES on Hiseq 2000 and Ion Proton, and low-depth WGS on Hiseq 2000) showed strong accordance. (B) Heat map showing amplifications (red) and deletions (blue) in 102 CCs. Samples are shown in rows. The diploid samples are clustered in the upper panel, and the multiploid samples are clustered in the lower panel. The proportion of genomic regions containing CNAs in each sample is shown in the right panel. (C) A sample with likely mutiploid genome which shown muti-region LOH and large segments with normal corpy-ratios and balanced allele frequencies. Copy-ratios (upper panel) and B allele frequencies (BAF,

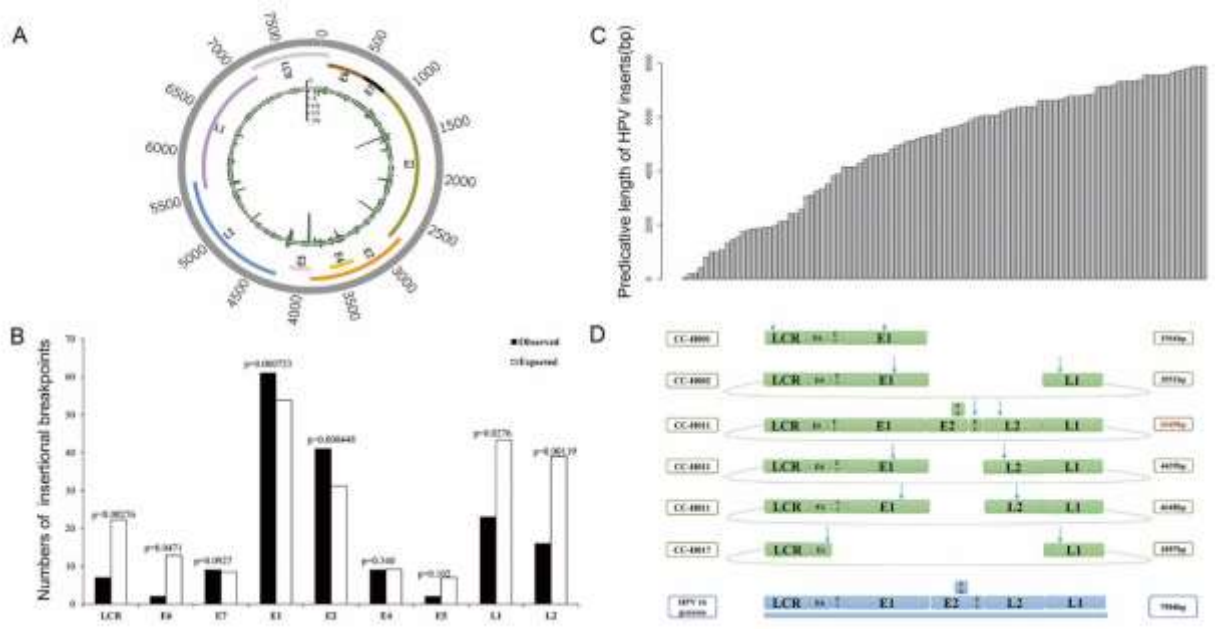
lower panel) are shown. (D) Significantly amplified (left) and deleted (right) regions were identified using GISTIC2.0. The total gene numbers and the potential driver genes for each region were annotated. (E) The structural variations in 20 CCs are shown in circos plots, with each plot representing one tumor.



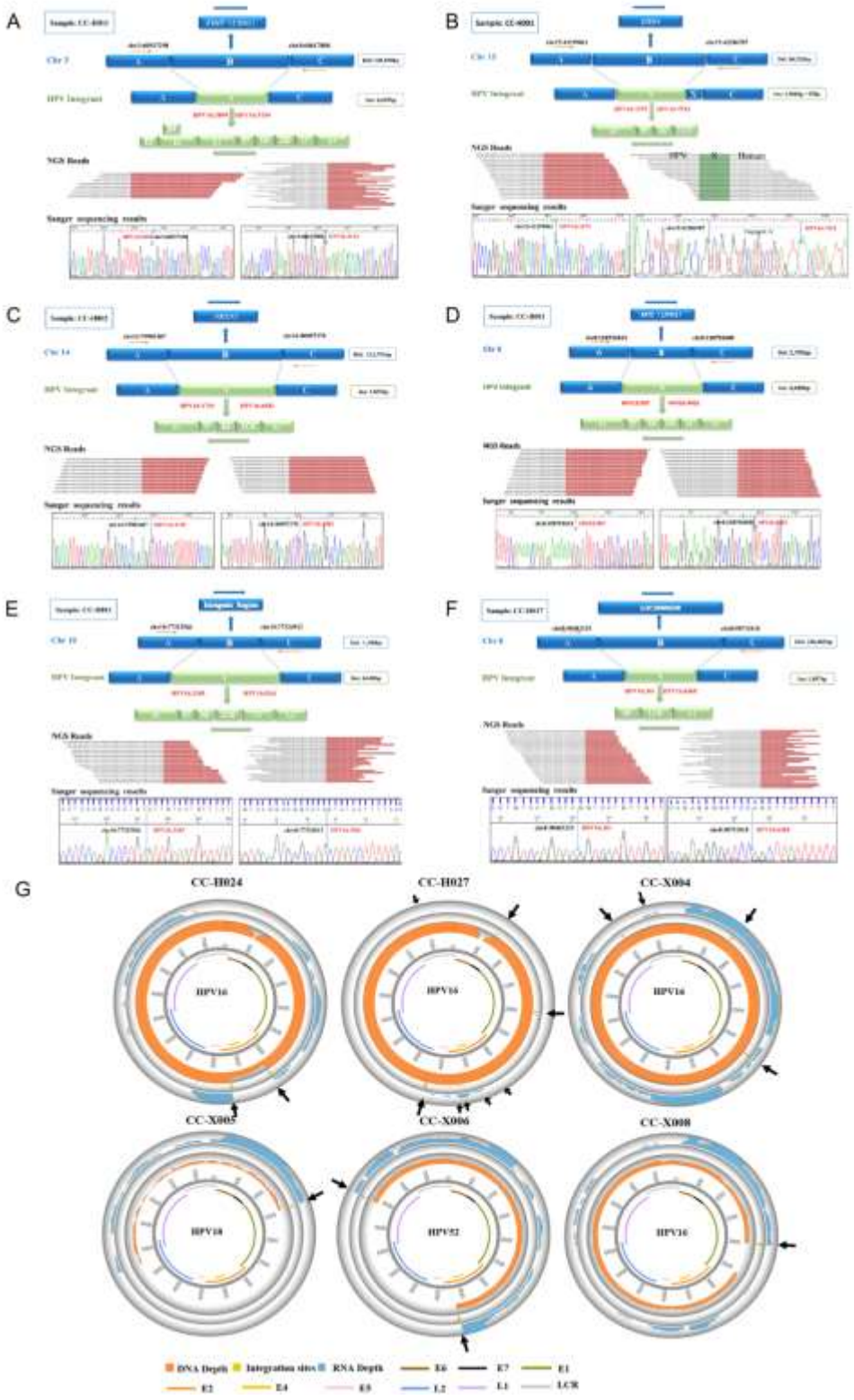
Supplementary Figure S4. Visualization of Mutated reads in Integrative Genomics Viewer. Four somatic mutations occurred in CIN specimens were shown as example. Only reads carrying mutations and with both mapping and base quality no less than 30 were shown.



Supplementary Figure S5. **ROC analysis of HPV testing as diagnostic marker for CC**. (A) ROC curve for detecting CC with the healthy using HPV integration or cytology-based testing. (B) Percentage of patients with HPV integration to those with HPV positive in different stages. Fisher's test was performed on the ratio of HPV integration between high-risk (HPV16, HPV18, HPV33, HPV58 plus other high-risk HPV types) and low-risk HPV types in different stages of HPV-positive participates.

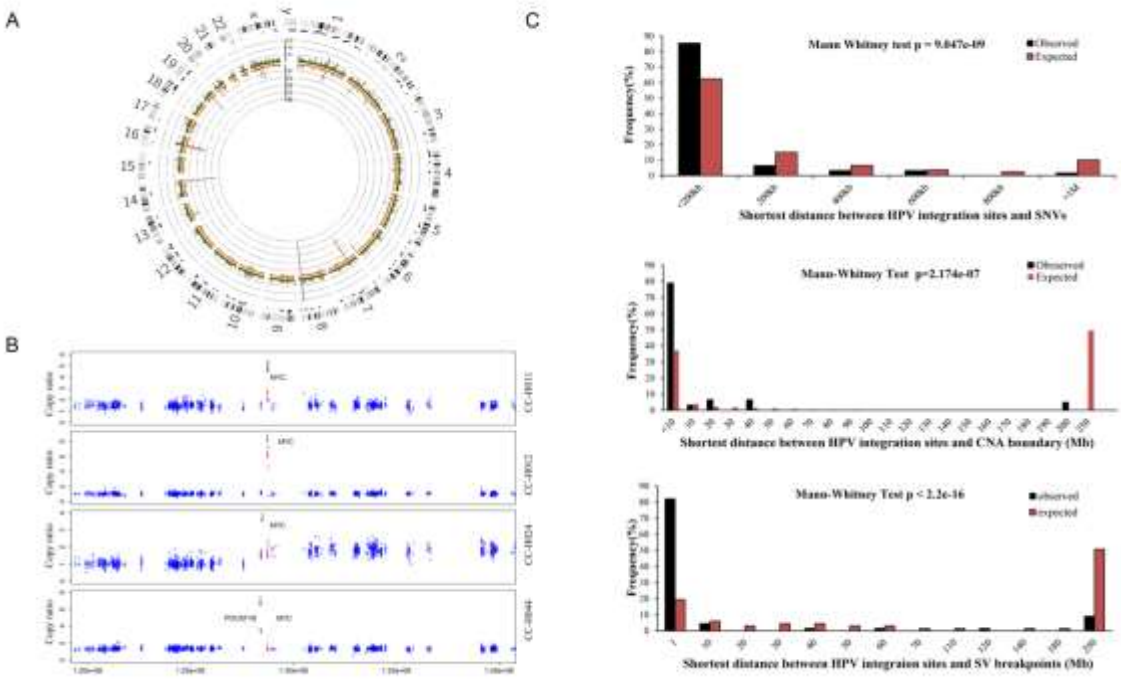


Supplementary Figure S6. Insertion of HPV fragments into human genome. (A) Circos plots of the distribution of HPV integration breakpoints in the HPV16 genome. HPV genes and locations are shown in the outer circle, and the distribution of the integration breakpoints in each 20-bp bin are shown in the inner circle; the green bars indicate supported reads with integrations in each bin, and the gray bars indicate the number of samples showing integrations in that region. **(B)** Expected integration breakpoints (one million) were randomly generated and compared with the observed breakpoints. Statistical analysis of the distribution of the integration breakpoints in the HPV genome (χ^2 test). **(C)** The 99 HPV integration fragments predicted using 353 integration breakpoints (≥10 supporting reads) in the CCs are sorted by size. **(D)** Six integration fragments of HPV16 were validated using long PCR-based Sanger sequencing. The arrows indicate integration breakpoints.



Supplementary Figure S7. HPV integration and expression. (A-F) Six HPV integration events in the genome of CC-001, CC-002, CC-011 and CC-017 validated using long PCR-based Sanger sequencing. Deletions of human genome sequence and

insertions of HPV genome fragments were shown. Next-generation sequencing reads and Sanger sequencing data supported this result. **(G)** Circos plots showing the distribution of HPV integration breakpoints and HPV gene expression in six samples. The inner circle shows the genome structure of the HPV strains, and the outer lanes show the DNA reads (orange), integration breakpoints (yellow and black arrows) and RNA reads (blue, shown as log2(reads number)).



Supplementary Figure S8. HPV integration and genome instability in human genome. (A) Distribution of HPV integration sites in the human genome. Chromosomes (outer panel), fragile sites (black dots) and integration sites in each 1-Mb bin (inner panel) were shown. The length of the bar represents the sum of supporting reads with integration sites that occurred in each bin (orange) and carcinomas with integration sites in that bin (green). (B) Focal amplification and large region gains in MYC gene and adjacent regions. (C) The distances between SNVs,

CNAs and SVs and the nearest HPV integration site in the human genome were analyzed and are shown in histograms. Analyses were based on whole-genome sequencing data generated on the HiSeq X Ten platform. Expected integration breakpoints were randomly generated in human genome, with the number equal to the observed ones. Statistical analyses were performed using Mann-Whitney tests.