Supplementary information

1. Samples

1.1 Study populations

This study was conducted at Peking Union Medical College Hospital (PUMCH) with the approval of its institutional review board. All 1728 participants of this study were pathologically diagnosed with cervical carcinoma (**Supplementary Figure 13**). Genomic DNA samples (n = 191) extracted from the primary, metastatic, or recurrent tumor tissue from 121 patient were sequenced by whole-exome sequencing(WXS) (n = 139) and whole-genome sequencing(WGS) (n = 52)， Patients in the stage IA-IIB had undergone a radical hysterectomy as the initial treatment and postoperative radiotherapy, no preoperative chemotherapy (to exclude the genotoxic effect of chemotherapeutic drugs), and received postoperative concurrent chemotherapy. And patients in stage III-IV were treated with radiotherapy and concurrent chemotherapy. We dichotomized patients who were followed for more than 5 years to poor prognosis group (recurrence or metastasis) and good survival group. Tissues from the primary tumor and tumor-adjacent (as the reference for detecting somatic alterations if available) were collected before initial treatment. All recurrent and metastatic tumor tissues were collected after chemoradiotherapy. Patients' clinical data and sequencing sample acquisition were displayed in **Supplementary Table 1** and **Supplementary Table 3**.

1.2 Treatment

Radiation therapies were administrated following NCCN guidelines (version 1.2012) as external beam radiation therapy and intracavity brachytherapy. The external beam radiation therapy was delivered with intensity-modulated radiation therapy (IMRT). CT simulation imaging was used to delineate gross tumor volume (GTV) and clinical target volume (CTV). CTV included the gross tumor, gross tumor volume of lymph nodes(GTVnd), cervix, uterus, upper part of the vagina (vagina stump for postoperative patients), parametrium, and pelvic lymph node regions (including the common iliac, external iliac, obturator, internal iliac, and presacral lymph node regions), and superior border of abdominal aorta bifurcation. For standard IMRT, 50.4 Gy in 28 fractions were administered to the planning CTV, and 59 to 61 Gy were delivered to the planning GTVnd with the simultaneous integrated boost.

Intracavity brachytherapy was delivered with ^{192}Ir, 30 to 36Gy in 5 to 7 fractions to the point A, while 10Gy was prescribed to a depth of 0.5 cm to cover proximal 3–5 cm of the vagina stump in 303 patients who received radical surgery.

Concurrent chemotherapies were 4-6 cycles of cisplatin 40 mg/m^2 weekly for all patients.

1.3 Clinical data

Medical records at PUMCH were retrospectively reviewed for demographic data, clinical and histopathologic information, and treatment parameters. Dates of death, cancer recurrence, and metastasis were confirmed by either querying the medical records or making telephone interviews. After completion of the initial therapy, patients returned to the outpatient clinic every 3 months for 2 years and every 6 months thereafter. Gynecologic examination and vaginal cytology were conducted at each visit, and radiologic examinations, such as CT, MRI, and PET, were performed when clinically indicated. The follow-up period for analysis of survival data ended in October 2017. To avoid bias, we only included sequencing patients that were followed up for a minimum of 5 years or until the time of death, recurrence, or metastasis. Instead, we used disease-free survival (DFS) as the endpoint of the outcome. DFS was defined as the interval from the date of the initial surgery to the date of recurrence, metastasis, or death by cervical cancer or the final follow-up. We dichotomized patients' survival status to poor survival (recurrence, metastasis after chemoradiotherapy) and good survival (no recurrence or metastasis was found during 5-year follow-up). We performed clinical matched samples based on the following parameters: diagnosis age, FIGO stage, histology type, tumor size, lymph node metastasis (**Supplementary Table 2**). This approach identified approximately equal numbers of patients with good or poor survival who were well balanced in terms of clinic-pathological features, enabling unbiased comparison between the status of genomic variation and patient outcome. All patients were diagnosed by two expert gynecological pathologists and one expert gynecological pathologist carefully reviewed the specimens to confirm the pathological diagnosis and tumor contents. Only squamous cell carcinoma, adenocarcinoma, and adenosquamous carcinoma were included in this study

(**Supplementary Table 1**). Specimens in which tumor cells (> 50%) and necrosis (< 20%) of the whole tissue were processed for sequencing analysis. Genomic DNA quantity and quality were determined to further exclude unqualified specimens. Clinical data for each sample was listed in **Supplementary Table 1**.

1.4 Board spectrum cancer data

Samples (*n* = 3188) in broad-spectrum cancers were collected for validating from seven TCGA cohorts including CESC(Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma), ACC(Adrenocortical Carcinoma), BRCA(Breast Invasive Carcinoma), LUAD(Lung Adenocarcinoma), LUSC(Lung Squamous Cell Carcinoma), HNSC(Head and Neck Squamous Cell Carcinoma) and ESCA(Esophageal Carcinoma) from the NIH GDC(Genomic Data Commons) database. Among them, 391 patients met our definition of more than 5 years of follow-up with different prognoses after radiotherapy. Two key words ‘radiation_therapy’ and ‘additional_radiation_therapy’, in the clinical information table, were used to pick up the patients receiving the radiotherapy treatment. Samples with regional recurrence, progression, or distant metastasis that occurred after radiotherapy were classified as the group with poor prognosis. This processing ensured that the grouping criteria are consistent with ours.

The detailed patient information was listed in **Supplementary Table 4**.

Samples (*n* = 18,210) from Genomics Evidence Neoplasia Information Exchange (GENIE) database revealed the frequency of genomic deletions of FANCD2 and DNMT1 in 35 types of cancers.

2. Methods

2.1 Sample processing

Tumor and adjacent normal DNA were extracted from 10 sections with a thickness of 5μm from FFPE tissues using QIAamp DNA FFPE Tissue Kit (QIAGEN) in the discovery set. For the validation set, GeneRead DNA FFPE Kit (QIAGEN) was used to reduce the artificial mutation of

C>T caused by deamination1. Paired blood cell DNA was extracted following instructions using the QIAamp DNA Blood Mini Kit (QIAGEN). DNA was quantified by Qubit (Life Technologies) and DNA integrity was examined by Agilent 2100 Bioanalyzer system (Agilent Technologies) or agarose gel electrophoresis.

2.2 HPV detection by fPCR

HPV infection was determined by an ultra-sensitive method using fluorescence PCR (fPCR) with high-risk HPV typing DNA kits (Shanghai Zhijiang Biotechnology Co., Ltd). Multiplex PCR amplification was used to detect 16 high-risk HPV types (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73 and 90) and 2 low-risk HPV types (HPV 6 and 11). Each sample was repeated at least 3 times.

2.3 Whole-exome sequencing

A paired-end DNA library was constructed according to the manufacturer's instructions (Agilent). Genomic DNA from patients was sheared into 250-350bp fragments by Covaris S220 sonicator and purified using AMPure SPRI beads from Agencourt. The DNA fragments were enriched by 6 cycles of PCR using SureSelect Primer and SureSelect ILM Indexing Pre-Capture PCR Reverse Primer. The size distributions of the libraries were examined with Agilent Bioanalyzer DNA 1000 chip. 500 ng DNA was subjected to whole-exome capture, using Agilent's SureSelect Human All Exon V5 Kit. The captured DNA-RNA hybrids were recovered using Dynabeads MyOne Streptavidin T1 from Dynal. DNA was eluted from the beads and desalted using QIAGEN MinElute PCR purification columns. The purified capture products were then amplified using the SureSelect ILM Indexing Post Capture Forward PCR Primer and PCR Primer Index 1 through Index 16 (Agilent). 50 Mb of DNA sequences of 334,378 exons from 20,965 genes was captured. DNA libraries were sequenced on Illumina Hiseq 4000 and Hiseq X Ten sequencing platform (Illumina) according to the manufacturer's instructions for paired-end 150bp read (Novogene, Beijing). The targeted sequencing depth ranged from 200× to 400×.

2.4 Whole-genome sequencing

0.5 μg genomic DNA from each sample was used as input for the DNA library preparations. The sequencing library was prepared using Truseq Nano DNA HT Sample Prep Kit (Illumina) following the manufacturer's recommendations. Briefly, genomic DNA sample was fragmented, end-polished, A-tailed, and ligated with the full-length adaptor for Illumina sequencing, followed by further PCR amplification. The clustering of samples was performed on a Cluster Generation System using Hiseq PE Cluster Kit (Illumina) according to the manufacturer's instructions. After cluster generation, the DNA libraries were sequenced on Illumina Hiseq X Ten sequencing platform (Illumina) and 150bp paired-end reads were generated (Novogene, Beijing). The sequencing depth was 40×.

2.5 Read alignment

Data were aligned to GRCH38/hg38 (ftp://hgdownload.cse.ucsc.edu:/goldenPath/hg38) + 143 HPV virus sequences obtained from HPVdetector[1] with bwa v0.7.17[2]. Defaults were used in bwa mem with the exception that four threads were utilized (-t 4) and shorter split hits marked as secondary (-M). This output SAM file was converted to a BAM file by using the Samtools v0.1.19 (https://github.com/samtools/samtools/), name sorted (samtools sort -n), mate pairings assigned (samtools fixmate), resorted by position (samtools sort), and indexed using 'samtools index'.

2.6 Read duplication marking and merging

Duplicate reads from the same sequencing library were merged using Picard v2.0.1 (https://github.com/broadinstitute/picard/) MergeSamFiles and duplicates were then marked per library using Picard MarkDuplicates. Each per-library BAM with marked duplicates was merged to generate a single BAM file for each sample. For both tools, 'ASSUME_SORTED=TRUE' and 'VALIDATION_STRINGENCY=SILENT' were specified. All other parameters were set to defaults.

2.7 Somatic mutation calling

Somatic point mutations and InDels were detected using GATK v3.7.0 with MuTect2 (-dontUseSoftClippedBases -max_alt_alleles_in_normal_count 2 -max_alt_alleles_in_normal_qscore_sum 40 -enable_clustered_read_position_filter -enable_strand_artifact_filter -annotateNDA). Variants supported less than 3 different reads and marked with 'strand_artifact', 'homologous_mapping_event', 'str_contraction', 'germline_risk'and 'multi_event_alt_allele_in_normal' were filtered.

2.8 Calibrate somatic muatation in WXS samples

As GeneRead DNA FFPE Kit was not used in the discovery set, there were false-positive C>T mutations caused by deoxygenation. A random forest (RF) model was built to calibrate somatic mutations. 19 features 'TLOD', 'AF_T', 'AD_T',' MEDIAN_LEFT_OFFSET', 'TLOD_FWD', 'TLOD_REV', 'TUMOR_SB_POWER_REV', 'MEDIAN_RIGHT_OFFSET', 'NLOD', 'TUMOR_SB_POWER_FWD', 'ND_T', 'AD_N', 'MAD_RIGHT_OFFSET', 'ND_N', ' MAD_LEFT_OFFSET', 'AF_N', 'FOXOG_T', 'HCNT', and 'ENCT' from MuTect2 results and one feature 'is C>T|G>A ?' for each site were obtained as model input. Training set samples came from three paired frozen vs FFPE tumor samples (SRP065941) and 3 paired lymph nodes (LN) vs primary tumor samples in our study. Another 3 paired Lymph node vs primary tumor samples were used as the validation set. All somatic mutations in training and validation sets were called following the procedure described above. The overlap somatic sites from paired samples (frozen vs FFPE and LN vs primary) were used as positive. From the non-overlapping sites, the same number of sites as the positive set were randomly selected as the negative. The random forest (RF) model was training by using scikit-learn (v0.19.1) with 4-fold cross-validation. The accuracy of the model was evaluated by the validation set. Finally, the model was used for recalibrating all somatic sites called in the WXS samples.

2.9 Potential genetic site filtering

To exclude potential genetic sites from somatic mutations, two databases (dbsnp v151 and ExAc v.0.3) were collected, and a normal pool was built by using the public SRA data listed in Supplementary Table 5 with the method described in the GATK pipeline.

All putative somatic mutations were filtered with the following criterion:

1) In dbsnp: CAF (allele frequency in 1000Genomes) > 0.001 or SAO = 1 (germline variants) or TOPMED (TOPMed allele frequency) > 0.001
2) In ExAc: allele frequency > 0.001
3) Less than two allele count in the normal pool

2.10 Mutation annotation

All somatic mutations were annotated by Ensembl Variant Effect Predictor (VEP, release 92)[3] and COSMIC database (v85)[4]. Mutations in the coding exons of transcripts with a complete open reading frame, and at the canonical splice donor or splice acceptor were retained. Other intronic variants, intergenic variants, and variants in the 5' flanking region, 3' flanking region, 5'UTR, and 3'UTR were removed.

2.11 Identifying significantly mutated genes (SMGs)

Mutations were included for all sets of 121 patients. MutSigCV(v1.41)[5] was utilized to identify SMGs. All SMGs were analyzed using a *P*-value cutoff of 0.01, showed in Supplementary Table 8.

TCGA-CESC data set (**Supplementary Table 4**) and two cervix cancer samples (**Supplementary Table 6**) were also used for SMGs identification. Venn analysis was performed in our dataset, TCGA-CESC dataset, and Meta dataset (Our PUMCH dataset, TCGA-CESC, and the two cervix cancer samples).

2.12 Mutation signature generation

All somatic point mutations were included to calculate relative weights of mutational signatures in a given sample. Non-negative matrix factorization (NMF) was used to statistically quantify the contribution of each signature for each tumor[6]. The NMF equation could be written as

$$\boldsymbol{X} = \boldsymbol{W} \times \boldsymbol{H} + \boldsymbol{U} \tag{1}$$

where $\boldsymbol{W}$ ($n\_samples \times n\_components$) and $\boldsymbol{H}$ ($n\_components \times n\_features$) are selected to optimally factorize or decompose the original mutation matrix $\boldsymbol{X}$.

Cluster analysis was carried on the cosine similarity of the signature ($\boldsymbol{H}$) obtained by our data decomposition, TCGA-CESC decomposition, and known signatures (downloaded from https://cancer.sanger.ac.uk/cosmic/signatures).

With the known signatures ($\boldsymbol{H}$) above, the signature scores ($\boldsymbol{W}$) for 121 sample were transformed by the following equation:

$$\boldsymbol{W} = \boldsymbol{X} \times \boldsymbol{H'} \times (\boldsymbol{H} \times \boldsymbol{H'})^{-1} \tag{2}$$

Then, Wilcoxon rank-sum test was used for signature scores among the good and poor survival group.

2.13 Identification of HPV integration in NGS data

Aligned reads were filtered by the ‘Duplicated’ reads and ‘XA’ tag. The sample considered HPV-integration positive if there were at least 3 flanking paired reads (PE reads) and 3 split reads (SR reads) supporting an integration site. The precise PE and SR reads were identified by the lumpy(v0.2.13)[7] with the filtered bam files (the ‘Duplicate’ reads and ‘XA’ tag reads were removed).

2.14 Copy number alteration (CNA) analysis

For WXS data, two methods were used for CNV calling in our data.

1) Control-FREEC(v.11.3)[8] was utilized to computes, normalizes, segment copy number and beta allele frequency (BAF) profiles, then calls copy number alterations and loss of heterozygosity (LOH). The matched normal sample was used as a control if available. And

known SNP sites of Han Chinese in Beijing, China (CHB) and Southern Han Chinese (CHS) whose minor allele frequency (MAF) > 5% in the 1000 Genomes Project were used for BAF estimating.

2) GATK (v.4.0.3) somatic CNA was performed according to the GATK pipeline. Tumor adjacent tissue and CHB, CHS population data were used to create the CNV panel of normals (PoN).

Amplification and deletion regions were retained when two methods were confirmed. The intersection regions were extracted by using bedtools (v2.27.1)[9].

For WGS data, only Control-FREEC was utilized to compute, normalize, segment copy number and beta allele frequency (BAF) profiles, then to determine copy number alterations and LOH. The matched normal sample was used as a control if available and known SNP sites of CHB and CHS (MAF > 5%) in the 1000 Genomes were used for BAF estimation. The mappability data created by GEM library (release 3)[10] with 2 mismatches and read length 150bp, was adopted for whole-genome CNA calling. The following two kinds of CNAs with high false-positive were filtered:

1) CNAs overlapped with regions in blacklist derived from FFPE samples.
2) The deletion regions < 500kb or the amplification region < 2Mb.

2.15 Chromosomal instability analysis

Seven masked CNA datasets from the Cancer Genome Atlas (TCGA) (ACC, BRCA, LUSC, LUAD, HNSC, CESC, and ESCA), downloaded from the Genomic Data Commons (GDC) database[11], were used to analyze the relationship between FANCD2 loss (log segments ratio < -0.4) and chromosomal instability (CIN). To assess structural and numerical CIN, the weighted genome instability index (wGII)[12] was used, which estimates the proportion of the genome with aberrant copy number compared with the median ploidy, weighted on a per-chromosome basis. The wGII was expressed as:

$$\mathbf{wGII} = \frac{\sum_c^m \sum_i^k l_i / L_c}{m} \tag{3}$$

where m was the number of chromosomes, k was the number of the CNA segments for each chromosome, L_c was chromosome length, and l_i was the length of the CNA segment.

2.16 CNA profile of tumor suppressor genes (TSGs)

The Tier 1 TSGs (n = 271) were collected from the Cancer Gene Census (CGC) database in COSMIC. With their CNA spectrum, samples in TCGA cohorts were hierarchical clustered using Euclidean distance and average linkage. The significance of TSG losses in FANCD2 deletion samples was analyzed by Fisher exact test and Wilcoxon rank test.

2.17 Mutual exclusivity of CNAs

Monte-Carlo permutation test (n = 10,000) was performed for mutual exclusivity of each deletion and amplification region. The p-value was calculated by:

$$p\text{-value} = 1 - \left(\sum_i^n M_i > M_{observe}\right)/n \tag{4}$$

where M is the number of co-occurrence of each deletion and amplification region in the sample set. The minus $\log_{10}$ transformed P value (scaled to 0~1) was used to measure the mutual exclusivity between the deletion and amplification regions.

2.18 Survival analysis

For the univariate model, considering a single mutated gene, Kaplan-Meier survival analysis (with log-rank test) was used to display the difference between the groups. For the multivariate and regression model, a Cox proportional-hazards regression model was applied to assess whether mutated-genes or signature scores were associated with DFS. All survival analyses were performed by using lifelines package (v0.14.6) with python (v2.7.3).

2.19 Tumor subpopulations identification and evolution

CNAs information derived from FREEC, as well as somatic mutations and SNPs in areas of CNAs, were combined to infer the tumor purity and phylogenies using package 'EXPANDS'[13] with the R language (v3.5.0). The function 'runExPANdS' set with default parameters except maxScore, which was lowered to 2.5 to reduce the false-positive rate of subpopulation detection. Subpopulations with size (cellular frequency) larger than 0.1 were considered. Mutations that could not be assigned to a high confidence subpopulation were discarded so that no ambiguous assignments were made. Mutations are assigned to all nested subpopulations (i.e. if a mutation is found in a subpopulation of cells at a high frequency, it will also be assigned to "daughter" subpopulations).

Phylogenetic relationships between the subpopulations inferred by the EXPANDS algorithm in all tumors per patient were generated using both somatic mutations and copy number alternations. The 'buildMultiSamplePhylo' function was used to calculate pairwise distances between subpopulations, and the 'bionjs' method was performed to generate phylogenies. The visualization of the evolution tree was performed by the ETE toolkit (v3.1.1) with python (v2.7.3).

3. Genomic analysis

3.1 Quality control of somatic mutations in FFPE samples

False-positive mutations of C>T/G>A were common in FFPE samples[14-16], therefore we used the 'GeneRead DNA FFPE Kit' for genomic DNA extraction for WGS, which effectively reduced DNA oxidation and limited false-positive somatic mutations[17]. For WXS, the traditional DNA extraction reagent 'QIAamp DNA FFPE Tissue Kit' was used, resulting in the percentage of artificial C>T/G>A mutations in WXS were much higher than TCGA-CESC cohort and the WGS data (**Supplementary Figure 14A**).

To reduce the false-positive somatic mutations, a random forest model was built from the 19 features described in the "Methods" section, and all of the features were derived from the GATK-Mutect2 outputs. The training set included two parts: 1) The raw sequence data were collected from BioProject SRP065941, the somatic mutations identified in both fresh-frozen (FF) and FFPE from the same sample were added into the positive set, while the rest somatic mutations in FFPE were used as negative set. 2) Somatic mutations called from FFPE samples of three patients, including both primary tumor and lymph node metastasis data. The shared somatic mutations in both primary and lymph node metastasis from the same patient were identified as the positive mutations, and the rest mutations were classified as negative ones. Then, we trained a random forest model using scikit-learn package (v0.19.1) in Python (v2.7.3) and achieved an accuracy of 95% in 4-fold cross-validation (**Supplementary Figure 14B**). The accuracy in a testing set has reached 85%, consisting of another 3 patients with paired primary tumor and lymph node metastasis data. Finally, all somatic mutations in WXS data were calibrated with this model. A large proportion of the artificial C>T/G>A mutations have been effectively removed (Supplementary Figure 14A). 34.0% of the calibrated mutations recurred in the COSMIC database (v85) (**Supplementary Table 7**). The NMF signatures decomposed from calibrated mutation profiles were similar to the ones from COSMIC, TCGA-CESC, and WGS data (**Supplementary Figure 14D**). Among them, signatures 1 and 2, cataloged in COSMIC, could be found in all types of cancer and cervical cancer.

In the Supplementary Figure 14C, the feature importance analysis showed the 'TLOD' (Tumor "logarithm of the odds" score) and 'AF_T' (allelic frequency in the tumor) were the two most contributing features for the calibration model and were often used to filter somatic mutations in FFPE samples[15, 18, 19].

3.2 Quality control of CNAs in FFPE samples

False-positive CNAs were common in FFPE samples, especially for the GC abnormal or lower mappability regions[20]. Therefore, we filtered out the false-positive CNAs commonly found in FFPE samples according to the blacklist by Scheinin *et al.*[20].

We further analyzed FFPE samples ($n = 4$) from 2 patients with 2 identical tumor lesions to assess the reproducibility of CNAs. As shown in Supplementary Figure 15A, more than 90% of the deletions were reproducible, while only 50% of amplifications were reproducible (Supplementary Figure 15B); most non-repeatable deletion fragments were small (< 500kb, 95%), while non-repeatable amplification fragments were relatively large (< 2Mb, 95%). Therefore, we filtered out genomic deletions with < 500kb sizes and genomic amplifications with < 2Mb sizes in this study.

3.3 Mutation Landscape of Cervical Cancer

To investigate the mutation landscape of cervical cancer in Chinese patients and compare with the cohort of 192 cervical cancer patients characterized by TCGA[18, 21], we first carried out the clinical cohort of 121 cervical cancer patients following chemoradiotherapy treatment, including patients with good prognosis (*n* = 45), patients with recurrence (*n* = 43), and patients with metastasis (*n* = 45) (**Figure 1**). Ten significantly mutated genes (SMG) were identified using the MutSigCV1.41 algorithm (*P* < 0.01, **Supplementary Figure 1** and **Supplementary Table 8**, EP300, FBXW7, PIK3CA, USP47, JKAMP, PTEN, PAK3, RB1, NFE2L2, and TP53. These SMGs are highly consistent with previously reported SMGs in the TCGA cohort[18, 21], while JKAMP and PAK3 are unique to this study. Further analyses with publicly available cervical cancer data showed 5 SMGs (EP300, FBXW7, PIK3CA, PTEN, and NFE2L2) are shared in all three datasets (Supplementary Figure 2 and Supplementary Table 9). Among them, 4 genes were found to have the mutation hotspots, including R2317W of EP300, R465H, and R505 of FBXW7, E542K, and E545K of PIK3CA, and R34P of NFE2L2 (**Supplementary Figure 3**). Interestingly, EP300 mutation rarely occurred in HPV18 positive tumor samples both in our PUMCH and TCGA-CESC dataset (**Supplementary Figure 4**).

Survival analysis and log-rank test revealed that a four SMG set (USP47, TP53, NFE2L2, RB1) was significantly associated with disease-free survival ($P = 6.69 \times 10^{-4}$, Supplementary Figure 5A). The four SMG set is further enriched in patients with recurrence ($P = 5.71 \times 10^{-3}$, **Supplementary Figure 5B**).

3.4 Molecular Characterization of Copy Number Alterations

Both Control-FREEC (v.11.3) and GATK (v.4.0.3) somatic CNA analyses were used for WXS and WGS of 121 patients and revealed several common focal deletions (Chromosome 3p, 4p,11q,13p) and amplification (Chromosome 3q, 11q,17q and 19) (**Supplementary Figure 7**). To identify candidate genes associated with the common CNAs, we applied logical regression analysis and revealed that genomic deletions of FANCD2 and RB1 were significantly associated with poor prognosis (**Figure 2A**).

To validate the genomic deletions, we carried out WXS and WGS both to confirm the FANCD2 deletions, the results of validated FANCD2 deletions shown in both patients (**Supplementary Figure 8B-D**). FANCD2 genomic deletions were also validated in two separate tissue samples (Primary tumor and lymph node metastasis) from one patient (**Supplementary Figure 8B and 8C**).

Enrichment of FANCD2 losses (deletions or mutations) at Chromosome 3p25.3 was identified using a likelihood-ratio test[22] (**Supplementary Figure 8A**). WGS data also showed an additional two patients with recurrence carried FANCD2 mutations, while one recurrent patient carried amplification of FANCD2 (Supplementary Figure 8A). The results revealed frequent genomic deletions (11 out of 43) and mutations (3 out of 43) of FANCD2 in 32.6% patients with recurrence and 15% (5 out of 32) patients with metastasis, only 4% (2 out of 45) patients with good prognosis showed loss of FANCD2.

3.5 Prognosis associated mutation signatures

The main somatic mutation profile in the PUMCH data consisted of seven known signatures, including signatures 1,2,3,5,6,13,16. Based on the multivariate Cox proportional-hazards regression model, signature 3 was significantly associated with poor prognosis in the PUMCH clinical matched data (P = 0.008, Supplementary Figure 6), which were found in breast, ovarian, and pancreatic cancers, and associated with failure of DNA double-strand break repair by homologous recombination.

4. Data Availability

All data could be found at the website: https://github.com/zju3351689/paper-PUMCH.

5. Code Availability

All software and scripts were listed in the **Supplementary Table 14**.

References:

1. Chandrani P, Kulkarni V, Iyer P, *et al.* NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. Br J Cancer 2015;112(12):1958-65.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754-60.
3. McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. Genome Biol 2016;17(1):122.
4. Forbes SA, Beare D, Boutselakis H, *et al.* COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45(D1):D777-D783.
5. Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499(7457):214-218.
6. Nik-Zainal S, Alexandrov LB, Wedge DC, *et al.* Mutational processes molding the genomes of 21 breast cancers. Cell 2012;149(5):979-93.
7. Layer RM, Chiang C, Quinlan AR, *et al.* LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014;15(6):R84.
8. Boeva V, Popova T, Bleakley K, *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28(3):423-5.
9. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841-2.
10. Derrien T, Estelle J, Marco Sola S, *et al.* Fast computation and applications of genome mappability. PLoS One 2012;7(1):e30377.
11. Grossman RL, Heath AP, Ferretti V, *et al.* Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 2016;375(12):1109-12.
12. Burrell RA, McClelland SE, Endesfelder D, *et al.* Replication stress links structural and numerical cancer chromosomal instability. Nature 2013;494(7438):492-496.
13. Andor N, Harness JV, Muller S, *et al.* EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics 2014;30(1):50-60.
14. Astolfi A, Urbini M, Indio V, *et al.* Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). BMC Genomics 2015;16:892.
15. Oh E, Choi YL, Kwon MJ, *et al.* Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. PLoS One 2015;10(12):e0144162.
16. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem 2015;61(1):64-71.
17. Bonnet E, Moutet ML, Baulard C, *et al.* Performance comparison of three DNA extraction kits on human whole-exome data from formalin-fixed paraffin-embedded normal and tumor samples. PLoS One 2018;13(4):e0195471.
18. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, *et al.* Integrated genomic and molecular characterization of cervical cancer. Nature

2017;543(7645):378-384.
19. Yost SE, Smith EN, Schwab RB*, et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Nucleic Acids Res 2012;40(14):e107.
20. Scheinin I, Sie D, Bengtsson H*, et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res 2014;24(12):2022-32.
21. Ojesina AI, Lichtenstein L, Freeman SS*, et al.* Landscape of genomic alterations in cervical carcinomas. Nature 2014;506(7488):371-5.
22. Cook JP, Mahajan A, Morris AP. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. Eur J Hum Genet 2017;25(2):240-245.