AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Deep Sequencing of HPV16 E6 Region Reveals Unique Mutation Pattern of HPV16 and Predicts Cervical Cancer

Wenchao Ai,[a] Chuanyong Wu,[b] Liqing Jia,[c] Xiao Xiao,[b] Xuewen Xu,[b] Min Ren,[c] Tian Xue,[c] Xiaoyan Zhou,[c] Ying Wang,[a] Chunfang Gao[a,b]

[a]Department of Laboratory Medicine, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai, China
[b]Clinical Laboratory Medicine Center, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China
[c]Department of Pathology, Fudan University Shanghai Cancer Center, Fudan University, Shanghai, China

Wenchao Ai, Chuanyong Wu, and Liqing Jia contributed equally to this work. Author order was determined on the basis of contribution.

ABSTRACT The genetic diversity of human papillomavirus (HPV) 16 within cervical cells and tissue is usually associated with persistent virus infection and precancerous lesions. To explore the HPV16 mutation patterns contributing to the cervical cancer (CC) progression, a total of 199 DNA samples from HPV16-positive cervical specimens were collected and divided into high-grade squamous intraepithelial lesion (HSIL) and the non-HSIL(NHSIL) groups. The HPV16 E6 region (nt 7125-7566) was sequenced using next-generation sequencing. Based on HPV16 E6 amino acid mutation features selected by Lasso algorithm, four machine learning approaches were used to establish HSIL prediction models. The receiver operating characteristic was used to evaluate the model performance in both training and validation cohorts. Western blot was used to detect the degradation of p53 by the E6 variants. Based on the 13 significant mutation features, the logistic regression (LR) model demonstrated the best predictive performance in the training cohort (AUC = 0.944, 95% CI: 0.913–0.976), and also achieved a high discriminative ability in the independent validation cohort (AUC = 0.802, 95% CI: 0.601–1.000). Among these features, the E6 D32E and H85Y variants have higher ability to degrade p53 compared to the E6 wildtype ($P < 0.05$). In conclusion, our study provides evidence for the first time that HPV16 E6 sequences contain vital mutation features in predicting HSIL. Moreover, the D32E and H85Y variants of E6 exhibited a significantly higher ability to degrade p53, which may play a vital role in the development of CC.

IMPORTANCE The study provides evidence for the first time that HPV16 E6 sequences contain vital mutation features in predicting the high-grade squamous intraepithelial lesion and can reduce even more unneeded colposcopies without a loss of sensitivity to detect cervical cancer. Moreover, the D32E and H85Y variants of E6 exhibited a significantly higher ability to degrade p53, which may play a vital role in the development of cervical cancer.

KEYWORDS cervical cancer, human papillomavirus type 16, next-generation sequencing, machine learning, E6 oncoprotein

HPV-induced cancers, particularly cervical cancer (CC), are expected to remain a significant global health challenge for the following decades (1). CC is the sixth most common cancer in Chinese women, with the development of 109,741 new cases and 59,060 deaths reported annually (2). It has been widely shown that persistent high-risk human papillomavirus (HPV) infections are the main risk factor for developing CC and its precursor lesions (3). As the most deleterious type, HPV16 was the leading cause of more than half of CC cases worldwide and has been detected in 50–55% of invasive CC cases (4, 5).

Editor Heba H. Mostafa, Johns Hopkins Hospital

Address correspondence to Ying Wang, nadger_wang@139.com, or Chunfang Gao, gaocf1115@163.com.

The authors declare no conflict of interest.

Received 23 April 2022
Accepted 1 June 2022
Published 23 June 2022

The direct colposcopy referral to of all women with positive results for the HPV16/18 genotyping test has been broadly suggested as a triage strategy of HPV-based screening programs (6). However, the major limitation of HPV16/18 genotyping tests is that we are unable to differentiate between transient and persistent infections. More than 90% of HPV16/18 infections will be cleared automatically by their innate immunity within a few years (7, 8). Just a limited proportion of patients with persistent infection of high-risk genotypes are at risk of developing tumors. Therefore, in order to reduce the burden of colposcopy referrals and related complications, a classification strategy is critical for patients who do not need colposcopy (9).

Multiple studies have revealed that the HPV variant has significant differences in the risk of HPV persistent infection and progression to cervical intraepithelial neoplasia (CIN) and CC (10–12). Meanwhile, several studies have analyzed the effect of HPV16 E6 oncoprotein variants overexpression in primary cultures of keratinocytes and found differences in their ability to induce serum/calcium resistant colonies and downregulation of p53/Bax (13), affecting several important cellular processes, including differentiation, apoptosis (14, 15), immortalization, transformation (16, 17), migration, and metastasis (17). Several studies have focused on the characterization of HPV16 gene variants in clinical samples obtained by traditional sanger sequencing methods (18–21). However, compared with the high sensitivity and specificity of the next-generation sequencing (NGS) (22), this method cannot accurately describe the complex sequence patterns of virus. The identification between HPV16-related high-grade squamous intraepithelial lesion (HSIL) and non-HSIL (NHSIL) based on the HPV16 E6 region using big data of NGS technology is still unexplored.

Previously, we have proposed a method to predict hepatocellular carcinoma using machine learning models based on HBV rt/s gene pattern features derived from NGS data (23, 24). In this study, we aimed to detect the E6 region of HPV16 positive patients by NGS to obtain the amino acid (aa) mutation features for individual HSIL prediction. Besides, to assess the carcinogenicity of the virus, we evaluate the differences in the degradation of p53 by HPV16 E6 mutations obtained from NGS.

RESULTS

Different HPV16 E6 feature characteristics between HSIL and NHSIL patients. HPV16 E6 mutation features were compared between HSIL and NHSIL patients based on the amino acid (aa) sequence of the HPV16 E6 gene (aa1–147) (Fig. 1A). The result showed that mutation features of the NHSIL patients were higher than those of HSIL patients in the E6 fragment, especially in the zinc finger (aa37–73 and aa110–146) regions (Fig. 1A). The cluster heatmap results were consistent with the circus plot. The E6 mutation features were different between HSIL and NHSIL patients, and the mutation frequency of HSIL patients was lower than that of NHSIL patients in general (Fig. 1B). We found that only 12 amino acid sites (1M, 5R, 8M, 12P, 17R, 27Q, 32D, 36E, 37C, 100N, 133H, and 144M) have a higher average mutation frequency in HSIL patients (Table S2 in the supplemental material). Meanwhile, five significantly different mutation sites (D32E, H85Y, L90V, Q98K, and R131K) between HSIL and NHSIL were identified using the unpaired Wilcoxon test and Fisher's exact test (Fig. 1 and Table S3). The above results indicated that great discrepancies were observed in mutation feature patterns between the HSIL and NHSIL patients.

Screening significant mutation features for model construction. When the missense mutation features of HPV16 E6 in HSIL and NHSIL patients were compared, a total of 33 differential mutation features were found (both Wilcoxon and Fisher's exact tests $P < 0.05$). Subsequently, we used the Lasso regression algorithm to select the most significant mutation features for classifying HSIL and NHSIL at the minimum value of the misclassification error using the cross-validation method (Fig. 2A and B). In total, 13 mutation features were filtered. The mutation features obtained by Lasso exhibited very different patterns between HSIL and NHSIL (Fig. 2C).

HSIL status prediction using machine learning approaches. Based on these 13 significant features by the Lasso algorithm, we then constructed the prediction models with the four machine learning approaches: logistic regression (LR), random forest (RF), support vector machine (SVM), and K-nearest neighbor (KNN). The area under the ROC curves (AUC) were

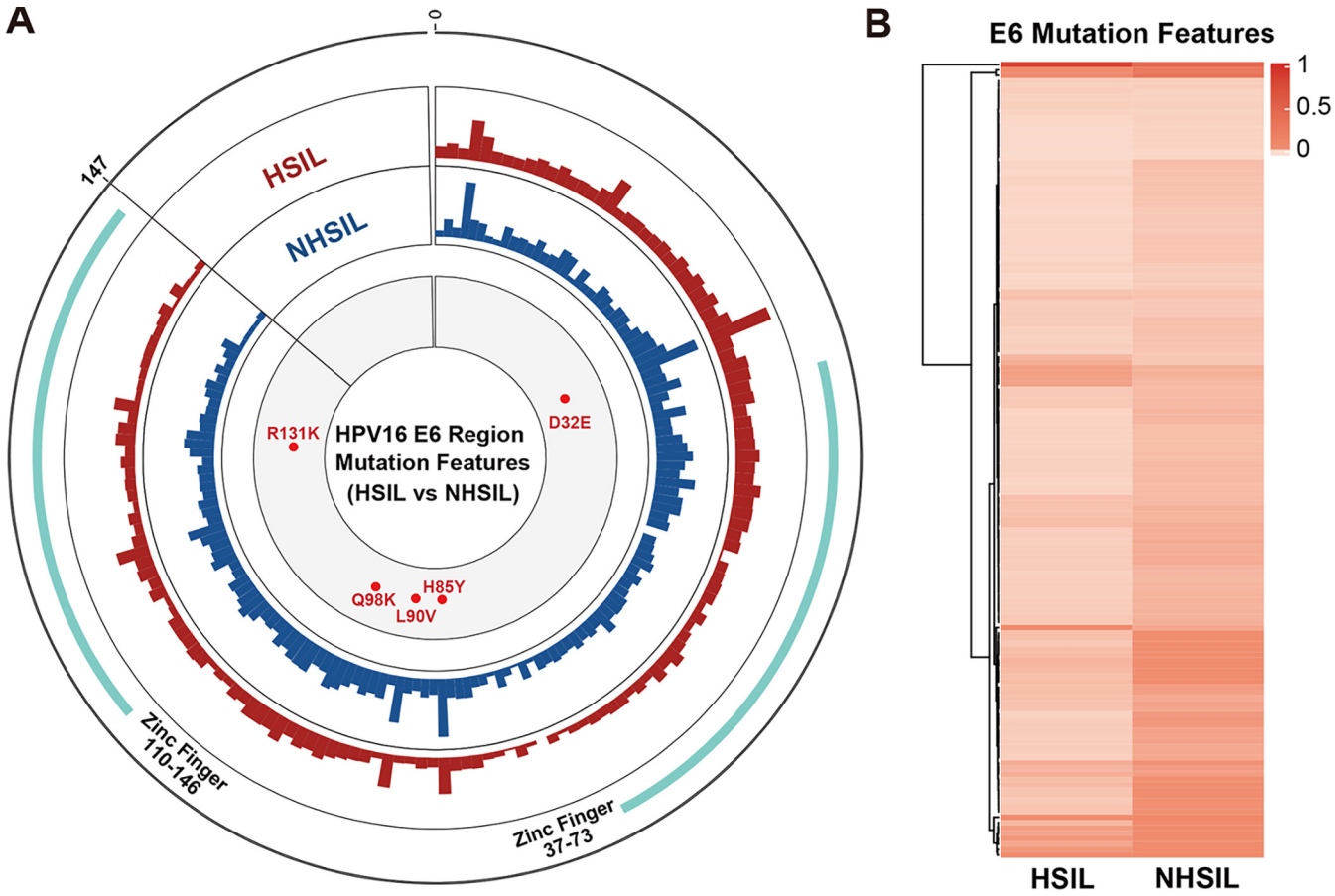

FIG 1 The complexity of the HPV16 E6 region between HSIL and NHSIL patients. (A) The outermost circle represents the location of amino acids (aa) of the HPV16 E6 protein (aa1–147). The important functional HPV16 E6 region distribution is marked as light green and as Zinc Finger (aa37–73 and aa110–146) (41). The colored histograms in the second and third circles indicate the complexity of mutation features (red, HSIL; blue, NHSIL) for each amino acid. The histogram represents the mutation features value. The innermost circle shows the differential aa mutations (red) between HSIL and NHSIL patients. (B) Different HPV16 E6 region mutation features among HSIL and NHSIL patients by hierarchical clustering. The value in the clustering map represents the mutation features of each group. A correlation was used for the sample measurement, Manhattan distance for the feature measurement, and ward.D2 algorithms for the clustering method.

used to evaluate their performance through 5-fold cross-validation in the training cohort (HSIL: $n = 123$ and NHSIL: $n = 43$). As shown in Fig. 3A, we found that the mutation features could accurately classify HSIL and NHSIL. It was worth noting that the LR model had a discriminating power of AUC values based on the mutation features in the training cohort (AUC = 0.944, 95% CI: 0.913–0.976) (Fig. 3A). For the independent validation cohort (HSIL: $n = 27$ and NHSIL: $n = 6$), the LR model also showed a satisfactory prediction performance (AUC = 0.802, 95% CI: 0.601–1.000) (Fig. 3B). The specificity and sensitivity of the LR model in differentiating HSIL from NHSIL were 83.3% and 81.9% in the independent validation cohort, respectively (Table 1). Thus, the LR algorithm based on the E6 mutation features achieved the best prediction performance in both training and validation cohorts.

The 3D structural analysis of HPV16 E6 mutations. Further analysis on the 13 mutation features identified by Lasso algorithm revealed the significant differences in D32E, H85Y, L90V, and Q98K mutations between HSIL and NHSIL patients. Previous studies have shown that HPV16 oncoprotein E6 binds to a short Leucine-rich LxxLL consensus sequence within ubiquitin ligase E6AP, resulting in the E6/E6AP heterodimer leading to p53 degradation (25). In this study, the D32E mutations were adjacent to the E6-p53 interface in the HPV16 E6/E6AP/p53 core ternary complex (Fig. 4A), while the H85Y, L90V, Q98K, and R131K mutations were found to be located near the E6-E6AP interface (Fig. 4B), which may affect the E6-p53 interaction.

HPV16 E6 D32E and H85Y mutations drive enhanced degradation of p53. To analyze the effects of HPV16 E6 variants on cell phenotype and p53 degradation, we constructed

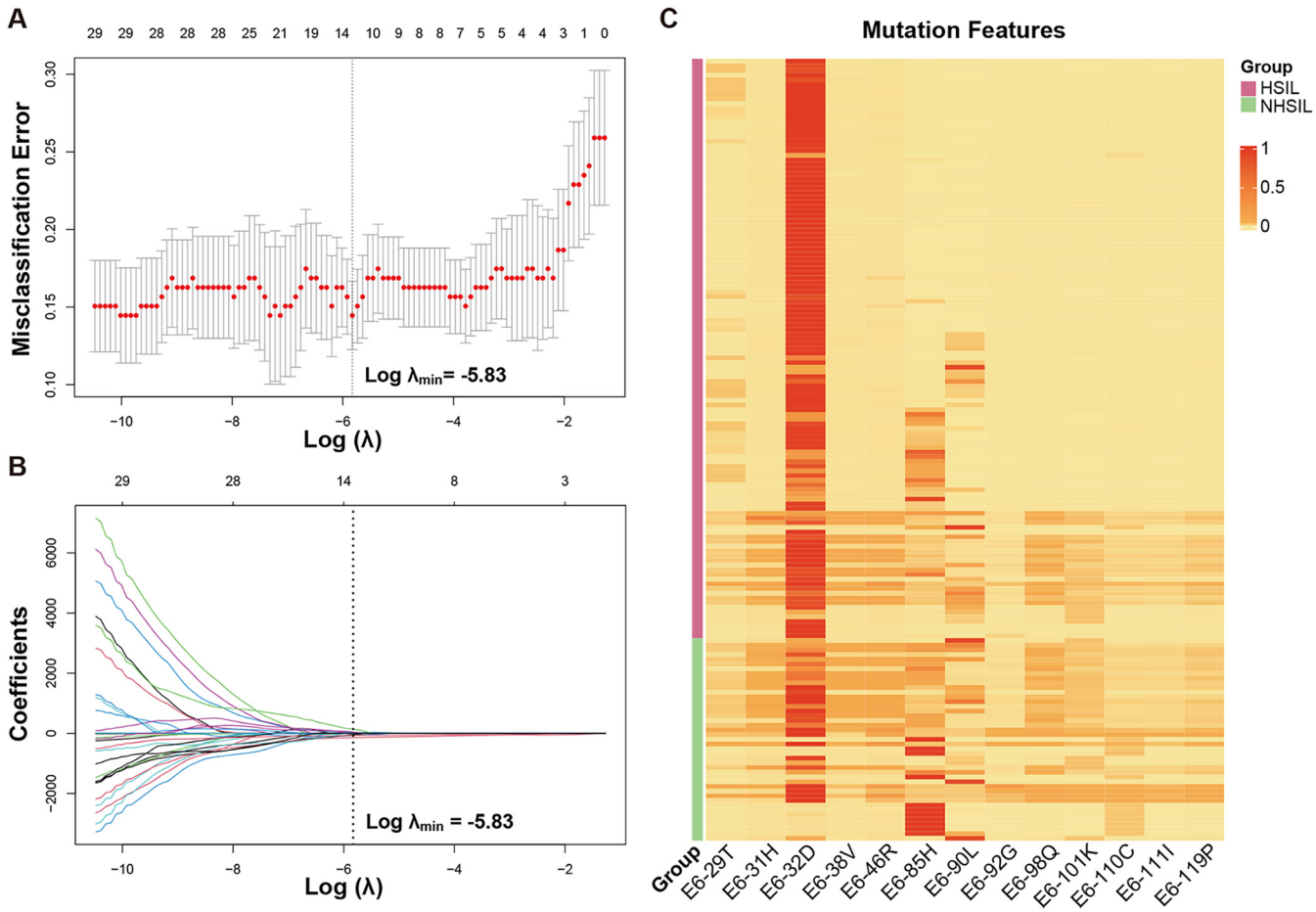

FIG 2 The dimensionality reduction results of HPV16 E6 mutation features by LASSO. (A, B) The coefficients in the Lasso regression for E6 mutation features screening. (A) For determining the adjustment parameter (λ value), we selected the adjustment parameter (λ value) with the minimum value of the misclassification error using the cross-validation method. (B) The adjustment parameter (λ value) was used to obtain the fraction deviance explained value. The obtained fraction deviance explained value was used to determine the final model selection. (C) Different mutation features among HSIL and NHSIL patients by hierarchical clustering.

plasmids of the E6 wild type (NP_041325.1) and E6 variants (D32E and H85Y) and expressed them in HPV-negative cervical cancer cell line. The cell proliferation assay showed that the D32E and H85Y variants had no effect on the cell proliferation (Fig. 5B). However, the Western blot results showed that both E6 wild type and E6 variants (D32E and H85Y) could degrade the p53 protein (Fig. 5C and D). And the D32E and H85Y variants exhibited a significantly higher ability to degrade p53 compared to the E6 wildtype ($P < 0.05$). Moreover, the H85Y variant was slightly more efficient in degrading p53 than the D32E variant ($P < 0.01$).

DISCUSSION

High-risk HPV persistent infection is well known to be a major risk of cervical cancer, with 70% of persistent infections attributed to HPV16 and 18 (26). As the major oncoprotein of HPV, E6 is related to the cellular immortalization, malignant transformation, and carcinogenesis through its abilities to degrade p53 (17, 25). To explore the vital features in HPV16 DNA sequences that can be used for CC risk prediction and act as a further triage for HPV-based genotyping tests, we propose a machine learning-based algorithm to distinguish HSIL patients from NHSIL patients based on HPV16 E6 sequence features derived from NGS.

In this study, we comprehensively discuss the mutation characteristics of the HPV16 E6 region in HSIL and NHSIL patients based on NGS. The result showed that the HSIL patients exhibited very distinct mutation patterns at the E6 amino acid level compared to the NHSIL patients. For women with positive results for HPV16 genotyping, the NGS-based HPV16 E6 sequence analysis demonstrated better discrimination of related progressive infections, which could reduce the colposcopy referral burden and associated complications.

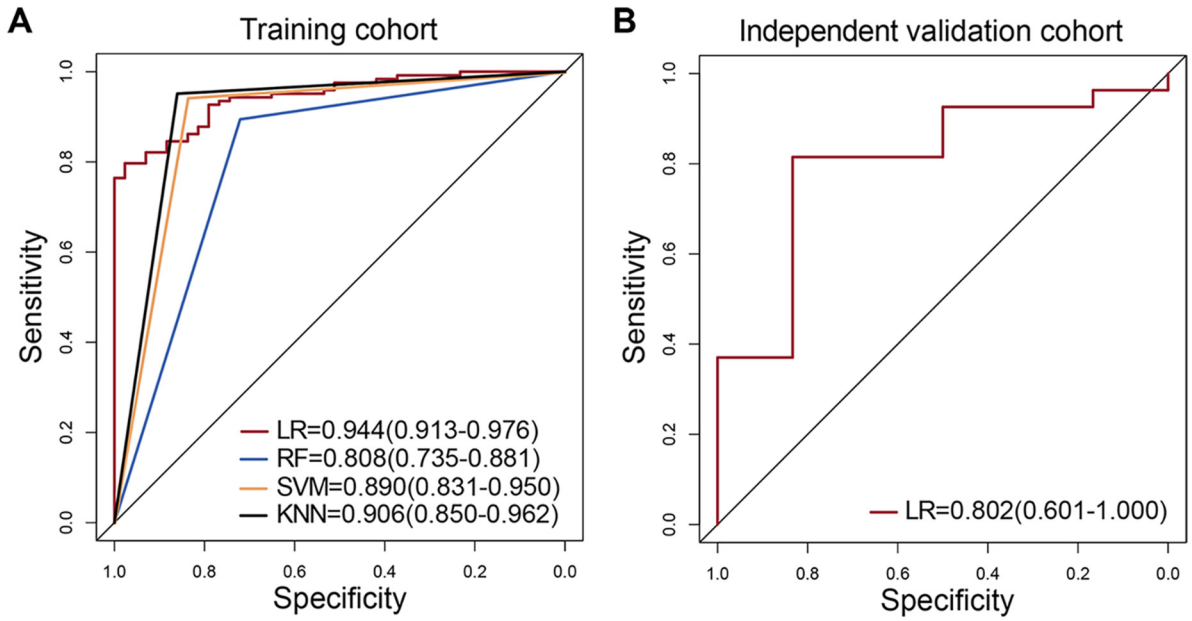

FIG 3 Four model performance in the training and validation cohorts. (A) Comparison of the AUC values between 4 models based on risk scores (red, LR; blue, RF; yellow, SVM; dark, KNN) in the training cohorts. (B) The AUC values of the LR model in the independent validation cohorts. The 95% confidence interval of AUC is also shown in the legend area.

The results showed that the mutation frequency of HSIL patients was generally lower than that of NHSIL patients (Fig. 1A), which is consistent with the findings of the previous study (27). The more complicated NHSIL mutations may be induced, at least in part, by the antiviral activity of human APOBEC3 (hA3) cytidine deaminases (28). It has been demonstrated that hA3A-mediated cytidine deaminase activity is capable of inducing HPV16 mutations (29) and suppressing HPV infectivity (30). Warren et al. (30) also proposed that these induced mutations, if not fatal, may be the cause of the long-term accumulation of genomic changes that lead to HPV-associated cancers.

In addition, to verify whether the subgroups of HSIL and NHSIL patients have an impact on the differences in E6 mutation features, we investigated the E6 mutation features among these subgroups (Fig. S3). Six higher frequencies of missense mutations including R62I, L90V, Q98K, E120D, Q130R, and R131K were identified in the subgroups of CC compared to that of CIN2/3 (Table S4, both Wilcoxon and Fisher's exact tests $P < 0.05$). However, only the E120D missense mutations showed significant difference between the subgroups low-grade squamous intraepithelial lesion (LSIL) and nonneoplastic/no evidence of disease (NED) (Table S4, $P < 0.05$). Between the subgroups of HSIL patients, the mutation features of zinc finger (aa37–73 and aa110–146) regions were higher in CIN2/3 than those in CC patients. However, this difference was not shown in the subgroups of NHSIL patients (Fig. S3). Our results showed that the more complex NHSIL mutations were indeed closely related to the low frequency mutation in HSIL patients (Fig. S3). In a previous study, Mirabello et al. (1) also evaluated rare variants in controls stratified according to cytological findings and found that HPV16-infected women with normal (NILM) cytology had significantly more variation than those

TABLE 1 Prediction performance of the machine learning models in the training and validation cohorts[a]

Cohort	Model	AUC (95% CI)	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV (%)	NPV (%)
Training cohort	LR	0.944 (0.913–0.976)	84.3	97.7	79.7	99.0	62.7
	RF	0.808 (0.735–0.881)	84.9	72.1	89.4	90.1	70.5
	SVM	0.890 (0.831–0.950)	91.6	83.7	94.1	94.3	83.7
	KNN	0.906 (0.850–0.962)	92.8	86.1	95.1	95.1	86.1
Independent validation cohort	LR	0.802 (0.601–1.000)	81.8	83.3	81.9	95.7	50.0

[a]LR, logistic regression; RF, random forest; SVM, support vector machine; KNN, K-nearest neighbor; AUC, area under curves; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

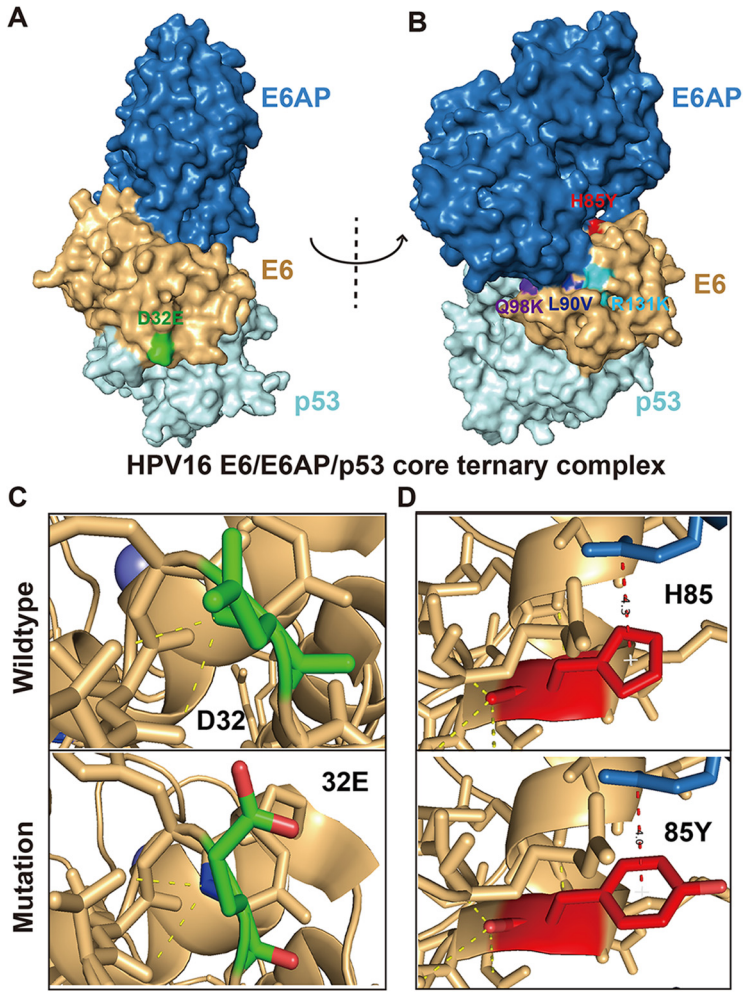

FIG 4 The 3D structural analysis of HPV16 E6 mutations. (A, B) Structure of the HPV16 E6/E6AP/p53 core ternary complex (25) (blue, E6AP peptide; brown, HPV16 E6; light cyan, p53 core). The mutations are marked on the surface by different colors (green, D32E; red, H85Y; dark blue, L90V; purple, Q98K; powder blue, R131K). (C) The 3D prediction structures of E6–p53 interface associated with D32E mutation. (D) The 3D prediction structures of E6–E6AP interface associated with H85Y mutation.

with cytomorphologic manifestations of an HPV infection (LSIL, $P = 0.01$), suggesting that variability might be related to decreased fitness.

The different E6 mutation feature patterns prompted us to explore the feasibility of using these different features of HPV16 E6 sequences to predict the risk of HSIL. Therefore, four different machine learning models were applied in this study, including LR, RF, SVM, and KNN. Evaluation of the AUC values showed that the LR algorithm achieved an approving diagnostic performance in both the training [AUC = 0.944 (0.913–0.976)] and the independent validation cohort [AUC = 0.802 (0.601–1.000)], indicating the potential clinical applications for HSIL status prediction (Table 1).

Recent studies have shown that novel molecular assays for recognizing proliferating cells and methylated target host genes could be used to triage HPV-positive cases (31). The Mexico cervical cancer screening study trial showed that in women with positive results for HPV16/18 genotyping, the p16/K_i-67 dual-stained cytology test performs better than cytology and E6 oncoprotein in discriminating relevant progressive infections (6). For the detection of CIN2+ lesions, the p16/K_i-67 dual-stained cytology was found to have higher sensitivity than liquid-based cytology (55.2% versus 23.9%) but slightly lower specificity (80.6% versus 87.5%) (6). Meanwhile, the p16/K_i-67 dual-stained cytology was found to have higher sensitivity than the E6 oncoprotein test (55.2% versus 31.3%) but somewhat lower specificity (80.6% versus 83.6%)

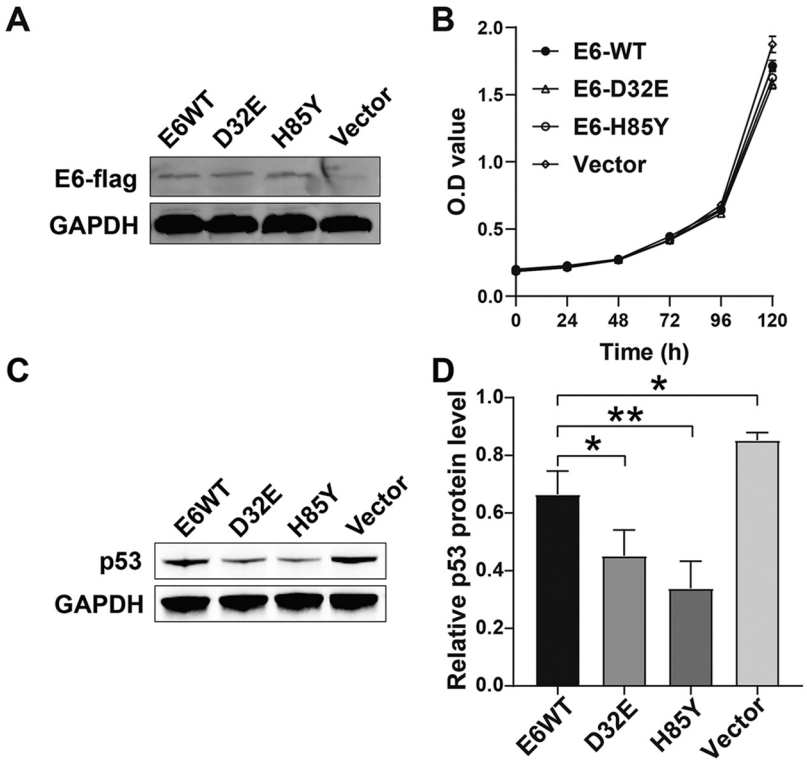

FIG 5 E6 variants (D32E and H85Y) drive enhanced degradation of p53. (A) E6-flag and GAPDH protein levels were assessed by Western blotting analysis. (B) CCK8 assay for exogenous HPV16 E6 expressing C-33A cells. (C, D) p53 and GAPDH protein levels were assessed by Western blotting analysis. GAPDH antibody was used as a loading control. *, $P < 0.05$; **, $P < 0.01$.

(6). p16/K_i-67 dual staining is effective, but it cannot be used to triage self-collected samples as reflex cytology is not possible (31).

The relationship between HPV16 E6 genome variations and the persistent viral infection and an increased risk of developing precancerous lesions and invasive CC had been widely reported and discussed. In France, Grodzki et al. (32) found that the HPV16 E6-350G variant presented an increased risk for development of cervical carcinoma, both for persistence (OR = 3.0; 95% CI: 1.4–6.7) and progression (OR = 6.2; 95% CI: 2.7–14.3). In Mexico, compared with the E-prototype, Ortiz-Ortiz (33) observed that AA-a showed a higher risk of developing CC. Mirabello (34) and Hang (35) found that the appearance of A4 variants among HPV16 positive women conferred a remarkably increased chance of gaining CC. These studies indicated that some HPV16-E6 variants might have a higher oncogenic potential.

Further analysis of the 13 mutation features identified by the Lasso algorithm revealed the significant differences in D32E, H85Y, L90V, and Q98K mutations between HSIL and NHSIL patients in our study. Except for D32E, the mutation frequency of other sites was higher in NHSIL patients (Table S3). Intriguingly, those E6 mutations were all located near the E6-E6AP or E6-p53 interface (Fig. 4), which might affect the E6-p53 interaction and ultimately cause p53 degradation.

As compared to the prototype, Hadami et al. (36) found that the highest p53 degradation was exhibited by the African variants Af2-a/r, Af1-d/G295, and Af2-a/G285, followed by the European variants E-C442/G350, E-G350/r, and NA1-b/r ($P < 0.05$). Moreover, Cuninghame et al. (37) found that the HIF-1α protein level and activity were increased by the Asian-American E6 (AAE6) variant through augment mitogen-activated protein kinase/extracellular related kinase signaling, leading to a hypoxia-tolerant phenotype with enhanced migratory potential. These studies suggest that HPV16-E6 mutation has biological effects on p53 degradation and metabolic reprogramming, which may play an important role in the occurrence and development of cervical cancer. However, in previous studies on the degradation of p53 by E6 variants, these variants often contain multiple mutation sites, which cannot accurately

evaluate the contribution of a single mutation site to the degradation of p53. In this study, we found that the D32E and H85Y variants exhibited a significantly higher ability to degrade p53 compared to the E6 wild type for first time ($P < 0.05$). However, it is noteworthy that, the same as the previous studies (18, 38), the mutation frequency of H85Y is higher in NHSIL patients, indicating that additional mechanisms may be involved in the epithelial transformation process.

Besides D32E and H85Y, we also identified some significant difference mutation sites including L90V, Q98K, and R131K between HSIL and NHSIL; the functions of these different missense mutations need to be further investigated. In addition, for more detailed estimating on the mutation rate of HPV16 and the accumulation of HPV16 variants in patients over time, we still need to conduct a longitudinal study in a larger sample of HPV16-infected populations.

In conclusion, the NGS-based HPV16 E6 sequence analysis in this study revealed more important genetic feature patterns between HSIL and NHSIL patients, highlighting the potential value of this test as a triage test in HPV screening programs. Combining HPV16 mutation patterns with NGS and machine learning methods may help to facilitate HSIL risk assessment and provide highly specific targets in etiology and treatment research. Moreover, to the best of our knowledge, this study is the first to find that the D32E and H85Y variants exhibited a significantly higher ability to degrade p53, which may play a vital role in the development of cervical cancer.

MATERIALS AND METHODS

Study population. From September 2020 to August 2021, a total of 199 DNA samples from HPV16-positive cervical specimens were collected and sequenced by NGS. Based on cytological and histological evaluations of fresh specimens, the cervical lesions were graded according to their severity as follows: 18 nonneoplastic/no evidence of disease (NED; chronic cervicitis and inflammation-related regenerative changes), 31 low-grade squamous intraepithelial lesion (LSIL) and CIN I, 29 CIN II/III, and 121 CC. The histological diagnosis of each case was reviewed by an experienced pathologist who was unaware of the HPV testing results. Histopathological findings were divided into a high-grade squamous intraepithelial lesion (HSIL, including CIN II/III and CC) group or a non-HSIL (NHSIL, including NED and LSIL) group.

Study design. The study was designed to reveal and assess the performance of the HPV16 E6 sequence features based on NGS in the identification and prediction of HSIL outcomes (Fig. 6) as follows.

Training phase: A total of 166 DNA samples from HPV16-positive cervical specimens were collected and divided into HSIL and NHSIL groups for model training. The aa mutation features of HPV16 E6 in HSIL and NHSIL patients were compared by Wilcoxon and Fisher's exact tests ($P < 0.05$). Then, the Lasso algorithm was used to select the most significant aa mutation features of HPV16 E6. Finally, the logistic regression (LR), random forest (RF), support vector machine (SVM), and K-nearest neighbor (KNN) were used to construct the diagnostic algorithms. The area under the receiver operating characteristics (AUC) curve was used to evaluate their diagnostic performance through 5-fold cross-validation.

Independent validation phase: In the independent validation cohort, 33 HP16-positive cervical specimens were collected to verify the algorithm with the best performance during the training phase.

Genomic DNA isolation, HPV typing, and sequencing. The supernatants were removed by centrifugation at 13,000 rpm for 10 min, and the pellets were collected for DNA extraction. Genomic DNA was extracted by nucleic acid extraction reagent (MCP-16C; Yaneng Biotechnology, Shenzhen, China). Human papillomavirus genotyping was conducted using a human papillomavirus genotyping Kit (Yaneng Biotechnology, Shenzhen, China).

After HPV testing, the remaining DNA samples were stored at −80°C. The HPV16 E6 region was amplified using Phusion High–Fidelity DNA polymerase (Thermo Fisher Scientific Baltics UAB, Vilnius, Lithuania) with specific primers. The following primers were used: 5′-AAACTAAGGGCGTAACCGAAATC-3′/5′-CAGCCTCTACATAAAACC ATCCAT-3′, and 5′-CAAGGAGACAGTTTATGCACCA-3′/5′-TGCAACAAGACATACATCGACC-3′. Subsequently, for each HPV16-DNA of samples, the HPV16 E6 region (nucleotides 7125–7566) sequencing was performed on the MiSeq sequencer with the MiSeq reagent kit, v3 (Illumina, San Diego, CA, USA) using Illumina paired-end sequencing protocols (Fig. S2). The procedures are described in the supplemental Materials and Methods.

Cell culture and transfection. The HPV-negative C-33A cell line was obtained from Cell Bank of the Chinese Academy of Sciences. The cell line was grown in MEM (Gibco by Invitrogen, Carlsbad, CA, USA) containing 10% fetal bovine serum (FBS; Gibco by Invitrogen, Carlsbad, CA, USA) and maintained at 37°C under humidified air and 5% CO^2. The cell line was authenticated using STR profiling within the last 3 years, and all experiments were performed with mycoplasma-free cells. The pcDNA3.1 vector carrying the HPV16-E6 wildtype (NP_041325.1), and the D32E and H85Y variants, were constructed by HANBIO (Shanghai, China). All plasmid transfections were conducted with Lipofectamine 3000 (Invitrogen, Life Technologies, Carlsbad, CA, USA) according to the manufacturer's instructions. The transfected cells were further cultured in medium supplemented with 500 μg/mL G418 (Solarbio, Beijing, China).

Cell proliferation assay. C-33A cell line transfected with empty pcDNA3.1 and pcDNA3.1 constructs containing the E6 wild type/variants were seeded in 96-well plates at a density of 2×10^4 cells per well. After 24, 48, 72, 96, and 120 h postseeding, 10 μL enhanced cell counting kit 8 (CCK8; Beyotime, Shanghai, China) was added to each well, and the cells were incubated for 2h at 37°C. The absorbance at 450 nm was measured using a microplate reader (Bio-Tek, VT, USA).

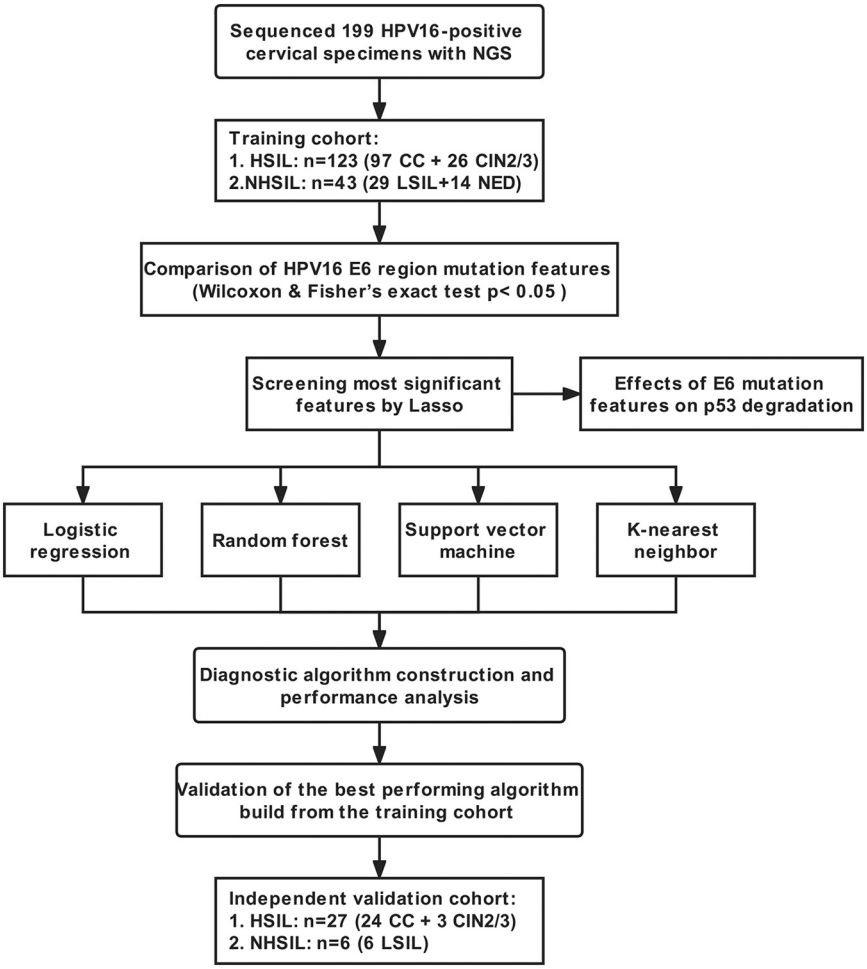

FIG 6 Study design and flowchart. A total of 166 DNA samples from HPV16-positive cervical specimens were collected and divided into HSIL and NHSIL groups for model training. In the independent validation cohort, 33 HPV16-positive cervical specimens were collected for subsequent analysis. CC, cervical cancer; CIN, cervical intraepithelial neoplasia; LSIL, low-grade squamous intraepithelial lesion; NED, nonneoplastic/no evidence of disease; LR, logistic regression; RF, random forest; SVM, support vector machine; KNN, K-nearest neighbor.

Western blots. Protein extraction was performed with a RIPA buffer mixture (Beyotime, Shanghai, China), and the proteins were then measured with a BCA protein assay kit (Solarbio, Beijing, China). The quantified samples were run on SDS-PAGE gels and were transferred onto PVDF membranes (Millipore, USA) using a transfer device. The membrane was blocked with 5% skim milk for 2h at room temperature, incubated with the primary antibody at 4°C overnight, and then probed with secondary antibody for 2 h. The target proteins were visualized by the Odyssey imaging system (LI-COR, USA). Antibodies specific for p53 (CST 2524), Flag (CST 8146), and GAPDH (Proteintech 60004-1) used for Western blot analysis were purchased from Cell Signaling Technology (Beverly, MA, USA) and Proteintech (Rosemount, IL, USA).

Data processing and analysis. Raw reads from a Miseq sequencer were processed by Cutadapt 3.5 to cut adaptor sequences and trim low-quality reads (base quality Q20). Filtered read pairs were aligned to the HPV16 reference genome sequence (accession number: NC_001526.4) using BWA 0.7.17.

The complexity of each aa position in the E6 region was calculated based on mutation features established by our previous research (23, 24). Mutations were identified and analyzed using R scripts. High mutation frequency was defined as a mutation rate of ≥5% of the total reads in each position. Statistical significance was evaluated using the unpaired Wilcoxon test and Fisher's exact test to identify differential mutations between the HSIL and NHSIL groups.

The mutation features of HPV16 E6 were shown by the BioCircos.js tool (39). The 3D structure of the HPV16 E6 protein was predicted based on PDB 4xr8 and 4yoz via PyMol 2.5.2 (25, 40). HPV16 E6 features models were trained to discriminate HSIL from NHSIL patients using the four machine learning approaches. The packages of "glmnet," "caret," "randomForest," "e1071," and "kknn" were conducted to calculate Lasso, LR, RF, SVM, and KNN, respectively, using R version 4.1.2 software. The predictive performance was measured by the receiver operating characteristic (ROC) curves. The ROC curve was used to calculate the optimal cutoff values that were determined by maximizing the Youden index. Accuracy of the optimal cutoff value was assessed by sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Ethics approval. The study was approved by the Institutional Ethics Committee of the leading medical center (Shanghai Eastern Hepatobiliary Surgery Hospital, EHBHKY2020-02-012). Written informed consent was obtained from all participants.

Data availability. The sequencing data used in this study have been submitted to NCBI under BioProject PRJNA830986.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.4 MB.

ACKNOWLEDGMENTS

Thanks go to Shanghai Amplicon-gene Bioscience Co., Ltd. and Shenzhen Yaneng Biotechnology Co., Ltd. for technical support.

This work was supported by the Innovation Group Project of Shanghai Municipal Health Commission [2019CXJQ03].

W. Ai and Y. Wang performed the methodology and data analysis and drafted the manuscript. C. Wu and L. Jia performed the experiments and provided clinical samples. X. Xu and X. Xiao collected the clinical information. M. Ren, T. Xue, and X. Zhou contributed to clinical testing and HPV typing. C. Gao got the funding, conceived and coordinated the overall study, and revised the manuscript.

The authors have declared that no competing interest exists.

All authors provided final approval of the manuscript and agree to be accountable for the accuracy and integrity of this work.

REFERENCES

1. Hoppe-Seyler K, Bossler F, Braun JA, Herrmann AL, Hoppe-Seyler F. 2018. The HPV E6/E7 oncogenes: key factors for viral carcinogenesis and therapeutic targets. Trends Microbiol 26:158–168. https://doi.org/10.1016/j.tim.2017.07.007.
2. Em FJ, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. 2020. Global Cancer Observatory: cancer today. International Agency for Research on Cancer, Lyon, France.
3. Münger K, Baldwin A, Edwards KM, Hayakawa H, Nguyen CL, Owens M, Grace M, Huh K. 2004. Mechanisms of human papillomavirus-induced oncogenesis. J Virol 78:11451–11460. https://doi.org/10.1128/JVI.78.21.11451-11460.2004.
4. Guan P, Howell-Jones R, Li N, Bruni L, de Sanjosé S, Franceschi S, Clifford GM. 2012. Human papillomavirus types in 115,789 HPV-positive women: a meta-analysis from cervical infection to cancer. Int J Cancer 131:2349–2359. https://doi.org/10.1002/ijc.27485.
5. Arbyn M, de Sanjosé S, Saraiya M, Sideri M, Palefsky J, Lacey C, Gillison M, Bruni L, Ronco G, Wentzensen N, Brotherton J, Qiao YL, Denny L, Bornstein J, Abramowitz L, Giuliano A, Tommasino M, Monsonego J. 2012. EUROGIN 2011 roadmap on prevention and treatment of HPV-related disease. Int J Cancer 131:1969–1982. https://doi.org/10.1002/ijc.27650.
6. Torres-Ibarra L, Lorincz AT, Wheeler CM, Cuzick J, Hernández-López R, Spiegelman D, León-Maldonado L, Rivera-Paredez B, Méndez-Hernández P, Lazcano-Ponce E, Salmerón J. 2021. Adjunctive testing by cytology, p16/Ki-67 dual-stained cytology or HPV16/18 E6 oncoprotein for the management of HPV16/18 screen-positive women. Int J Cancer 148:2264–2273. https://doi.org/10.1002/ijc.33414.
7. Rodríguez AC, Schiffman M, Herrero R, Wacholder S, Hildesheim A, Castle PE, Solomon D, Burk R, Proyecto Epidemiológico Guanacaste Group. 2008. Rapid clearance of human papillomavirus and implications for clinical focus on persistent infections. J Natl Cancer Inst 100:513–517. https://doi.org/10.1093/jnci/djn044.
8. Rodríguez AC, Schiffman M, Herrero R, Hildesheim A, Bratti C, Sherman ME, Solomon D, Guillén D, Alfaro M, Morales J, Hutchinson M, Katki H, Cheung L, Wacholder S, Burk RD. 2010. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. J Natl Cancer Inst 102:315–324. https://doi.org/10.1093/jnci/djq001.
9. Tota JE, Bentley J, Blake J, Coutlée F, Duggan MA, Ferenczy A, Franco EL, Fung-Kee-Fung M, Gotlieb W, Mayrand MH, McLachlin M, Murphy J, Ogilvie G, Ratnam S. 2017. Approaches for triaging women who test positive for human papillomavirus in cervical cancer screening. Prev Med 98:15–20. https://doi.org/10.1016/j.ypmed.2016.11.030.
10. Freitas LB, Chen Z, Muqui EF, Boldrini NA, Miranda AE, Spano LC, Burk RD. 2014. Human papillomavirus 16 non-European variants are preferentially associated with high-grade cervical lesions. PLoS One 9:e100746. https://doi.org/10.1371/journal.pone.0100746.
11. Cornet I, Gheit T, Iannacone MR, Vignat J, Sylla BS, Del Mistro A, Franceschi S, Tommasino M, Clifford GM. 2013. HPV16 genetic variation and the development of cervical cancer worldwide. Br J Cancer 108:240–244. https://doi.org/10.1038/bjc.2012.508.
12. Xi LF, Koutsky LA, Hildesheim A, Galloway DA, Wheeler CM, Winer RL, Ho J, Kiviat NB. 2007. Risk for high-grade cervical intraepithelial neoplasia associated with variants of human papillomavirus types 16 and 18. Cancer Epidemiol Biomarkers Prev 16:4–10. https://doi.org/10.1158/1055-9965.EPI-06-0670.
13. Asadurian Y, Kurilin H, Lichtig H, Jackman A, Gonen P, Tommasino M, Zehbe I, Sherman L. 2007. Activities of human papillomavirus 16 E6 natural variants in human keratinocytes. J Med Virol 79:1751–1760. https://doi.org/10.1002/jmv.20978.
14. Zehbe I, Richard C, DeCarlo CA, Shai A, Lambert PF, Lichtig H, Tommasino M, Sherman L. 2009. Human papillomavirus 16 E6 variants differ in their dysregulation of human keratinocyte differentiation and apoptosis. Virology 383:69–77. https://doi.org/10.1016/j.virol.2008.09.036.
15. Zehbe I, Lichtig H, Westerback A, Lambert PF, Tommasino M, Sherman L. 2011. Rare human papillomavirus 16 E6 variants reveal significant oncogenic potential. Mol Cancer 10:77. https://doi.org/10.1186/1476-4598-10-77.
16. Richard C, Lanner C, Naryzhny SN, Sherman L, Lee H, Lambert PF, Zehbe I. 2010. The immortalizing and transforming ability of two common human papillomavirus 16 E6 variants with different prevalences in cervical cancer. Oncogene 29:3435–3445. https://doi.org/10.1038/onc.2010.93.
17. Niccoli S, Abraham S, Richard C, Zehbe I. 2012. The Asian-American E6 variant protein of human papillomavirus 16 alone is sufficient to promote immortalization, transformation, and migration of primary human foreskin keratinocytes. J Virol 86:12384–12396. https://doi.org/10.1128/JVI.01512-12.
18. Zhao J, Zhu J, Guo J, Zhu T, Zhong J, Liu M, Ruan Y, Liao S, Li F. 2020. Genetic variability and functional implication of HPV16 from cervical intraepithelial neoplasia in Shanghai women. J Med Virol 92:372–381. https://doi.org/10.1002/jmv.25618.
19. Zhao JW, Zhan Q, Guo JH, Liu M, Ruan YT, Zhu TL, Han LF, Li F. 2019. Phylogeny and polymorphism in the E6 and E7 of human papillomavirus: alpha-9 (HPV16, 31, 33, 52, 58), alpha-5 (HPV51), alpha-6 (HPV53, 66), alpha-7 (HPV18, 39, 59, 68) and alpha-10 (HPV6, 44) in women from Shanghai. Infect Agents Cancer 14:11. https://doi.org/10.1186/s13027-019-0250-9.

20. Dai MZ, Qiu Y, Di XH, Shi WW, Xu HH. 2021. Association of cervical carcinogenesis risk with HPV16 E6 and E7 variants in the Taizhou area, China. BMC Cancer 21:769. https://doi.org/10.1186/s12885-021-08531-y.
21. Zhe XY, Xin HZ, Pan ZZ, Jin FY, Zheng WN, Li HT, Li DM, Cao DD, Li Y, Zhang CH, Fu SW, Shao RF, Pan ZM. 2019. Genetic variations in E6, E7 and the long control region of human papillomavirus type 16 among patients with cervical lesions in Xinjiang, China. Cancer Cell Int 19:11. https://doi.org/10.1186/s12935-019-0774-5.
22. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature 550:345–353. https://doi.org/10.1038/nature24286.
23. Chen S, Zhang Z, Wang Y, Fang M, Zhou J, Li Y, Dai E, Feng Z, Wang H, Yang Z, Li Y, Huang X, Jia J, Li S, Huang C, Tong L, Xiao X, He Y, Duan Y, Zhu S, Gao C. 2021. Using quasispecies patterns of hepatitis B virus to predict hepatocellular carcinoma with deep sequencing and machine learning. J Infect Dis 223:1887–1896. https://doi.org/10.1093/infdis/jiaa647.
24. Wang Y, Xiao X, Chen S, Huang C, Zhou J, Dai E, Li Y, Liu L, Huang X, Gao Z, Wu C, Fang M, Gao C. 2021. The impact of HBV quasispecies features on immune status in HBsAg+/HBsAb+ patients with HBV genotype C using next-generation sequencing. Front Immunol 12:775461–775461. https://doi.org/10.3389/fimmu.2021.775461.
25. Martinez-Zapien D, Ruiz FX, Poirson J, Mitschler A, Ramirez J, Forster A, Cousido-Siah A, Masson M, Vande Pol S, Podjarny A, Travé G, Zanier K. 2016. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. Nature 529:541–545. https://doi.org/10.1038/nature16481.
26. Forman D, de Martel C, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L, Vignat J, Ferlay J, Bray F, Plummer M, Franceschi S. 2012. Global burden of human papillomavirus and related diseases. Vaccine 30(Suppl 5):F12–F23. https://doi.org/10.1016/j.vaccine.2012.07.055.
27. Mirabello L, Yeager M, Yu K, Clifford GM, Xiao Y, Zhu B, Cullen M, Boland JF, Wentzensen N, Nelson CW, Raine-Bennett T, Chen Z, Bass S, Song L, Yang Q, Steinberg M, Burdett L, Dean M, Roberson D, Mitchell J, Lorey T, Franceschi S, Castle PE, Walker J, Zuna R, Kreimer AR, Beachler DC, Hildesheim A, Gonzalez P, Porras C, Burk RD, Schiffman M. 2017. HPV16 E7 genetic conservation is critical to carcinogenesis. Cell 170:1164–1174.e6. https://doi.org/10.1016/j.cell.2017.08.001.
28. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424:99–103. https://doi.org/10.1038/nature01709.
29. Vartanian JP, Guétard D, Henry M, Wain-Hobson S. 2008. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science 320:230–233. https://doi.org/10.1126/science.1153201.
30. Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D, Lambert PF, Santiago ML, Pyeon D. 2015. APOBEC3A functions as a restriction factor of human papillomavirus. J Virol 89:688–702. https://doi.org/10.1128/JVI.02383-14.
31. Bhatla N, Singhal S. 2020. Primary HPV screening for cervical cancer. Best Pract Res Clin Obstet Gynaecol 65:98–108. https://doi.org/10.1016/j.bpobgyn.2020.02.008.
32. Grodzki M, Besson G, Clavel C, Arslan A, Franceschi S, Birembaut P, Tommasino M, Zehbe I. 2006. Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350G variant. Cancer Epidemiol Biomarkers Prev 15:820–822. https://doi.org/10.1158/1055-9965.EPI-05-0864.
33. Ortiz-Ortiz J, Alarcón-Romero L. d C, Jiménez-López MA, Garzón-Barrientos VH, Calleja-Macías I, Barrera-Saldaña HA, Leyva-Vázquez MA, Illades-Aguiar B. 2015. Association of human papillomavirus 16 E6 variants with cervical carcinoma and precursor lesions in women from Southern Mexico. Virol J 12:29. https://doi.org/10.1186/s12985-015-0242-3.
34. Mirabello L, Yeager M, Cullen M, Boland JF, Chen Z, Wentzensen N, Zhang X, Yu K, Yang Q, Mitchell J, Roberson D, Bass S, Xiao Y, Burdett L, Raine-Bennett T, Lorey T, Castle PE, Burk RD, Schiffman M. 2016. HPV16 sublineage associations with histology-specific cancer risk using HPV whole-genome sequences in 3200 women. J Natl Cancer Inst 108:djw100. https://doi.org/10.1093/jnci/djw100.
35. Hang D, Yin Y, Han J, Jiang J, Ma H, Xie S, Feng X, Zhang K, Hu Z, Shen H, Clifford GM, Dai M, Li N. 2016. Analysis of human papillomavirus 16 variants and risk for cervical cancer in Chinese population. Virology 488:156–161. https://doi.org/10.1016/j.virol.2015.11.016.
36. Hadami K, Saby C, Dakka N, Collin G, Attaleb M, Khyatti M, Filali-Maltouf A, Morjani H, El Mzibri M. 2021. Degradation of p53 by HPV16-E6 variants isolated from cervical cancer specimens of Moroccan women. Gene 791:145709. https://doi.org/10.1016/j.gene.2021.145709.
37. Cuninghame S, Jackson R, Lees SJ, Zehbe I. 2017. Two common variants of human papillomavirus type 16 E6 differentially deregulate sugar metabolism and hypoxia signalling in permissive human keratinocytes. J Gen Virol 98:2310–2319. https://doi.org/10.1099/jgv.0.000905.
38. Wang Y, Tong Y, Zhang Z, Zheng R, Huang D, Yang J, Zong H, Tan F, Xie Y, Huang H, Zhang X. 2021. ViMIC: a database of human disease-related virus mutations, integration sites and *cis*-effects. Nucleic Acids Res 50:D918–D927.
39. Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, Yue H, Zhang P, Chen R. 2016. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. Bioinformatics 32:1740–1742. https://doi.org/10.1093/bioinformatics/btw041.
40. Guiley KZ, Liban TJ, Felthousen JG, Ramanan P, Litovchick L, Rubin SM. 2015. Structural mechanisms of DREAM complex assembly and regulation. Genes Dev 29:961–974. https://doi.org/10.1101/gad.257568.114.
41. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S, Redaschi N, Xenarios I, Bridge A. 2015. HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res 43:D1064–D1070. https://doi.org/10.1093/nar/gku1002.