Check for updates

Genome composition-based deep learning predicts oncogenic potential of HPVs

Lin Hao[1], Yu Jiang[2], Can Zhang[2] and Pengfei Han[2]*

[1]Department of Pharmacy, Linfen Central Hospital, Linfen, China, [2]The 4th Medical Center, People's Liberation Army (PLA) General Hospital, Beijing, China

Human papillomaviruses (HPVs) account for more than 30% of cancer cases, with definite identification of the oncogenic role of viral *E6* and *E7* genes. However, the identification of high-risk HPV genotypes has largely relied on lagged biological exploration and clinical observation, with types unclassified and oncogenicity unknown for many HPVs. In the present study, we retrieved and cleaned HPV sequence records with high quality and analyzed their genomic compositional traits of dinucleotide (DNT) and DNT representation (DCR) to overview the distribution difference among various types of HPVs. Then, a deep learning model was built to predict the oncogenic potential of all HPVs based on *E6* and *E7* genes. Our results showed that the main three groups of Alpha, Beta, and Gamma HPVs were clearly separated between/among types in the DCR trait for either *E6* or *E7* coding sequence (CDS) and were clustered within the same group. Moreover, the DCR data of either *E6* or *E7* were learnable with a convolutional neural network (CNN) model. Either CNN classifier predicted accurately the oncogenicity label of high and low oncogenic HPVs. In summary, the compositional traits of HPV oncogenicity-related genes *E6* and *E7* were much different between the high and low oncogenic HPVs, and the compositional trait of the DCR-based deep learning classifier predicted the oncogenic phenotype accurately of HPVs. The trained predictor in this study will facilitate the identification of HPV oncogenicity, particularly for those HPVs without clear genotype or phenotype.

1 Introduction

Human papillomaviruses (HPVs) are a group of double-stranded DNA (dsDNA) viruses, specifically tropic to human cutaneous and mucosal epithelial tissues of the anogenital tract, hands, or feet (Correa et al., 2017). There are more than 220 HPV types, which are classified into five genera—Alpha (65 genotypes), Beta (54 genotypes), Gamma (98 genotypes), Mu (four genotypes), and Nu (one genotype) (Muhr et al., 2018) (https://www.hpvcenter.se/)— according to the most genomically homologous viral gene of the major capsid *L1* gene.

HPVs exert significant oncogenic roles in various malignant types, unlike most commonly infected viruses, such as adenovirus (Liu et al., 2014), herpes simplex virus (Whitley and Roizman, 2001), influenza viruses (Sun et al., 2014; Deng et al., 2017), and coronaviruses (Sun et al., 2014; Deng et al., 2017). Infectious agents contribute to approximately 12% of global cancers annually (de Martel et al., 2020, 2017). Particularly, HPVs account for more than 30% of cases (https://gco.iarc.fr, accessed on November 15, 2023) (de Martel et al., 2020), such as squamous cell carcinoma (SCC) of the skin (Jablonska et al., 1972), cervical tumors (Walboomers et al., 1999), genital malignancies (de Villiers, 1994), anogenital malignancies (Zur, 2002), and head and neck cancers (Guo et al., 2018). Furthermore, the oncogenic roles of HPVs were heterogenized among various genotypes. HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, and HPV59 account for the most number of cancers of the cervix, anus, vulva, vagina, penis, and head and neck (Longworth and Laimins, 2004; Galati et al., 2024). HPV16 and HPV18, the most popular high-risk (HR) HPV strains (Bzhalava et al., 2013), account respectively for nearly 50% and 20% of cervical cancer cases (Li and Coffino, 1996; de Sanjose et al., 2010). Thus, it is important to predict and evaluate how an HPV promotes oncogenesis.

There are two main coding regions in the HPV genome (HPV16 as an example): the early (E) region encodes regulatory proteins of E6, E7, E1, E2, E4, and E5; the late (L) region encodes the structural L1 and L2 capsid proteins (Gheit, 2019). Moreover, E6 and E7 are significantly carcinogenic in various types of cancers (Barbosa, 1996; McLaughlin-Drubin and Munger, 2009) by regulating essential cellular progress. E6 protein promotes p53 degradation and thus antagonizes apoptosis (Martinez-Zapien et al., 2016; Shu et al., 2022). E6 also inhibits apoptosis by activating the transcription of survivin and the apoptosis inhibitor c-IAP2 (Bruyere et al., 2023) and regulating host antiviral responses (Yanatatsaneejit et al., 2020). Additionally, HR-HPV E6 binds the protein containing PDZ domains, facilitates its degradation, and thus promotes cell survival and proliferation (Accardi et al., 2011; Drews et al., 2019). The E7 protein also poses oncogenic roles by variously interacting with the DREAM protein complexes (Drosophila, RB, E2F, and Myb) (Rashid et al., 2015) and targeting the pRB family members, which is necessary for malignant transformation (Demers et al., 1996; Zur, 2000; Zhang et al., 2006). Furthermore, HPV E7 proteins exhibit conserved and virus type- and species-specific interactions with cellular proteins (Demers et al., 1996; Zur, 2000; Zhang et al., 2006). Therefore, it is vital to identify the exact association between the genotypes of HPV *E6/E7* and their phenotype for malignant transformation.

Prevalent identification of the association between the genotypes of HPV *E6/E7* and their malignant phenotypes has been recognized to depend on molecular biological and virological approaches (Li and Coffino, 1996; Accardi et al., 2011; White et al., 2012; Martinez-Zapien et al., 2016; Songock et al., 2017; Shu et al., 2022; Bruyere et al., 2023), implying an urgent need for intelligent and fast method to recognize or predict the benign or malignant phenotype of HPVs based on HPV *E6/E7* genotypes. Multiple artificial intelligence (AI) approaches have been utilized to represent genomic information of viruses and to predict viral phenotypes based on such represented information. The deep and hidden viral genotype–phenotype association has been well learned by machine learning (ML) or deep learning (DL) methods. The pandemic potential of influenza viruses (Li et al., 2020), the SARS-CoV-2 adaptation (Li et al., 2020), and the host adaptation of bat coronaviruses (Jiang et al., 2023; Li et al., 2023) have been accurately predicted by ML or DL models. Genomic information of amino acid (Babayan et al., 2018), dinucleotide (DNT) (Li et al., 2020), and DNT representation (DCR) (Li et al., 2022) of the viral genome is predictable and interpretable in the case of genotype–phenotype association. However, the oncogenic phenotype of HPVs analyzed using AI approaches has been less frequently reported.

In this study, we retrieved and cleaned HPV sequence records with high quality and analyzed their genomic traits such as DNT, DCR, and codon dependency. Then, we utilized unsupervised ML approaches to visualize the distribution difference among various types of HPVs. Finally, we built a deep learning model to predict the oncogenic potential of all HPVs based on *E6/E7* genes. The DL model trained in this study can predict the oncogenic potential of any HPV.

2 Methods

2.1 HPV genome sequence retrieval and data cleaning and the counting of compositional traits of HPV genes

Genome records for HPVs were retrieved using the keywords (("Human papillomavirus"[Organism] OR Human papillomavirus [All Fields]) AND complete [All Fields] AND genome [All Fields]) AND (biomol_genomic [PROP] AND ("6500" [SLEN]: "8500" [SLEN])) from the National Center for Biotechnology Information (NCBI) nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore). Records were cleaned with sequence quality by excluding the sequence without full coding sequence (CDS) of *E7*, *E6*, *E2*, *E1*, *L2*, and *L1* and the sequence with a ratio of undefined or degenerate bases of more than 2%. The oncogenic label for each sample was manually added according to reported research (Muhr et al., 2018). The compositional traits of DNT and DCR were counted according to the reported tools (Li et al., 2023, 2022). The *E6* and *E7* CDSs were parsed based on the "Location/Qualifiers" information for each record and then were counted under the reported guidance (https://github.com/Jamalijama/BatCoVadaptation). The counted DNT and DCR features were taken as a vector with fixed dimensions of 48 and 1,536. The DCR data were analyzed in detail and utilized for deep learning.

2.2 Data distribution analysis with unsupervised machine learning

HPV DCR data were analyzed for the distribution among various HPV types. The dimension reduction was performed using the reported python scripts by the method of t-distributed stochastic neighbor embedding (t-SNE) and principal component

analysis (PCA). The 1,536 features of the DCR trait for each sample were reduced into two main components (t-SNE1 and t-SNE2, or PCA1 and PCA2), with packages sklearn.decomposition.PCA and sklearn.manifold.TSNE, both of which were performed according to https://scikit-learn.org/stable/about.html#citing-scikit-learn. The reduced data were plotted for the data distribution difference using the Python Seaborn package of pairplot. Hierarchical clustering was performed to overview the clustering and separation of HPV samples with "type" annotation added into sequence accession. The clustering was calculated using the algorithm of Euclidean distance by the sns.clustermap package.

2.3 Building the deep learning model for oncogenic HPV prediction

The deep learning classifier for HPV oncogenicity was built using a model of convolutional neural network (CNN), according to previous reports (Li et al., 2023, 2022). The 1,536-dimensional DCR of each HPV sample (*E6* or *E7*), with an oncogenicity label of 1 or 0, was input into the model. Three rounds of convolution and pooling were performed to learn in-depth the hidden association of DCR with viral oncogenicity phenotype. HPV data were randomly split into training and validating sets using sklearn.model_selection import train_test_split with a validation data size of 25%. The 1,536-dimensional DCR was first reshaped into a matrix of $6 \times 16 \times 16$ for a 3D-CNN performance. Out-channels were set as 8, 16, and 32. A kernel size of (1, 3, 3), a rectified linear unit (ReLU) activation, and the average pooling were utilized for each CNN layer. Two linear transformations were performed to transform the 768-dimensional vector first into the 192-dimensional vector and then into the two-dimensional final predicted labels of 1 and 0. The Softmax function was utilized to output the prediction probability. A learning rate of 0.01 and a training epoch of 50 were utilized for either model.

To evaluate the prediction performance of each classifier, micro-average receiver operating characteristic (ROC) (Fawcett, 2005) and confusion matrix (Fawcett, 2005) were plotted. The full-connection data after the three rounds of convolution performance were reduced by PCA and then were plotted for the two main components.

2.4 Statistics

Statistical analysis was performed for the significance in the PCA1 value of the PCA-reduced full-connection data and analyzed using an unpaired, non-parametric Mann–Whitney test using the SciPy scripts of python.

3 Results

3.1 Pipeline of genomic trait analysis and oncogenic potential prediction of HPVs

Full HPV genome information with genomic DNA sequences and annotations was downloaded from the NCBI website. The 3,485

HPV records were cleaned to obtain high-quality samples. A total of 2,782 samples were parsed to obtain the annotation set of record ID, ORGANISM, and other information and then to obtain all CDS and protein sequences. Data were separated according to HPV type, and then Alpha HPVs with definite high or low oncogenicity were randomly split into training and testing sets. All other HPVs, particularly unclassified HPVs, were predicted using a trained deep learning predictor (Figure 1A). Genomic compositional traits of HPV *E6* and *E7* CDSs were counted according to the reported method (Li et al., 2023, 2022) (Figure 1B). Unsupervised ML approaches were performed to analyze the distribution difference of HPV genomic traits among various types. Features of DNT, codon, DCR, and others were first reduced using the method of PCA or t-SNE and then plotted for the two main components. Full feature data for each type of trait were clustered using a hierarchical clustering method (Figure 1C). A deep learning network of CNN was trained with the DCR trait data of HPV *E6* or *E7* for oncogenicity prediction of HPVs (Figure 1D). Finally, each HPV sample was predicted using a trained model, and the temporal/spatial distribution of HPVs with predicted high or low oncogenicity was analyzed (Figure 1E).

3.2 Data distribution of HPVs and type-based clustering and separation of HPVs revealed by unsupervised ML methods

The distribution of HPVs on the annotation of date, country, genotype, and isolation host was counted based on the parsed HPV data. Most of the recorded HPVs on BCBI were isolated post-2010 (Figure 2A). More than one-third of the samples were not annotated for their isolation country, and most of the annotated samples were from the USA, Luxembourg and the Netherlands (the EU), Japan and China (East Asia), and South Africa (Figure 2B). Approximately 1/10 HPV samples were not annotated for their isolation hosts, and the top genotypes were HPV16, HPV35, HPV52, HPV58, HPV18, HPV31, HPV11, HPV51, and HPV53, which are high or low oncogenic HPVs (Figure 2C). Additionally, the isolation annotation was also not clear for more than one-fifth of the samples (Figure 2D). Thus, it is important to build an intelligent predictor for identifying the oncogenicity of these HPVs with unknown annotations.

To overview the HPV distribution among the four types (Alpha, Beta, Gamma, and Mu) of HPVs or all genotypes, full data of compositional DCR or other traits were reduced into two main components with t-SNE or PCA and then plotted with type information annotated for each sample. As indicated, the main three groups (Alpha, Beta, Gamma, only two samples for Mu HPVs) were clearly separated in DCR trait for *E6* CDS (Figure 3A by t-SNE and Figure 3B by PCA) or *E7* CDS (Figure 3C by t-SNE and Figure 3D by PCA). Moreover, the Alpha HPVs were further analyzed using the same methods to visualize the distribution variation between high and low oncogenic HPVs. Both *E6* CDS (Figure 3E by t-SNE and Figure 3F by PCA) and *E7* CDS sequence (Figure 3G by t-SNE and Figure 3H by PCA) were clearly separated.
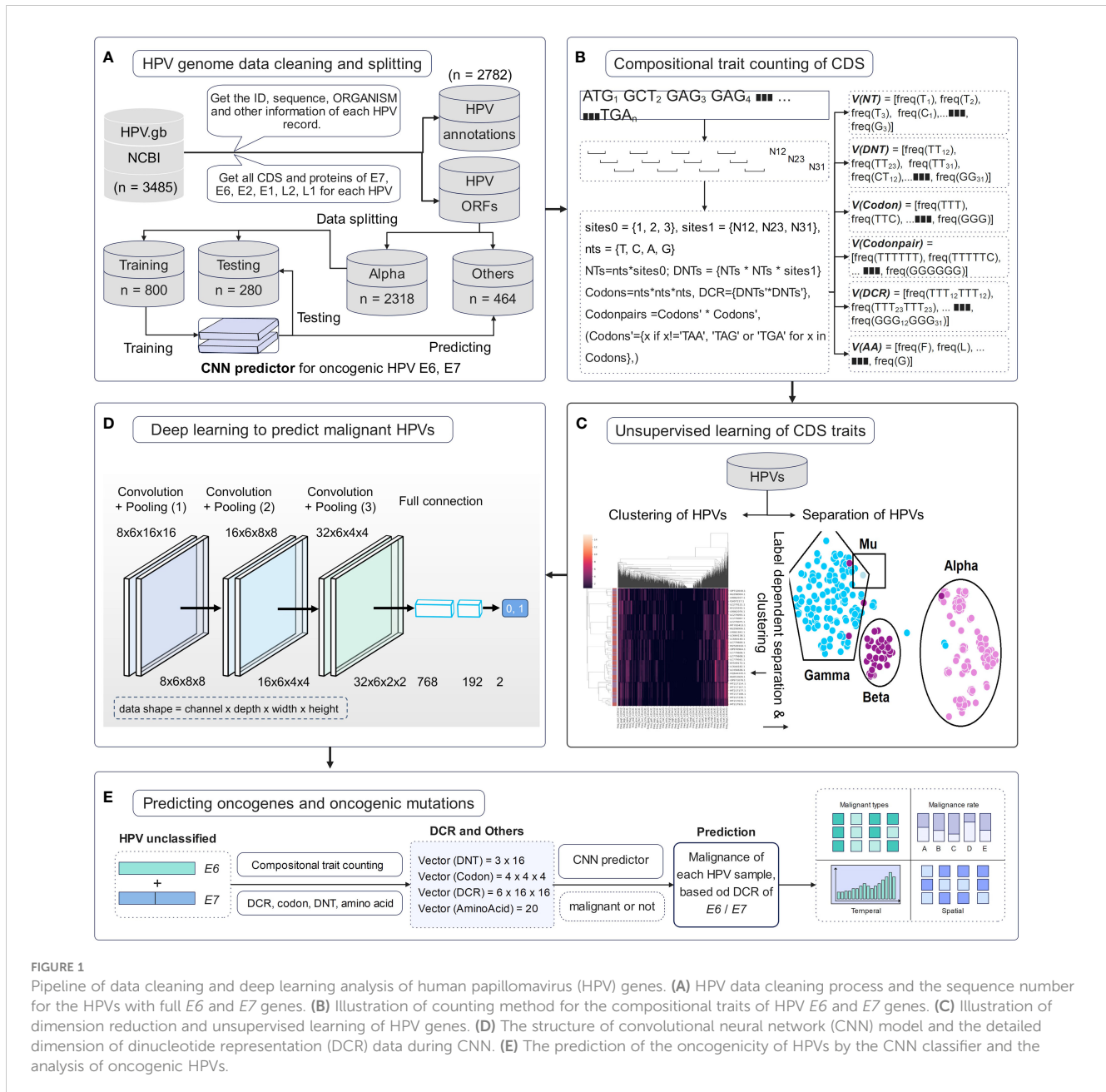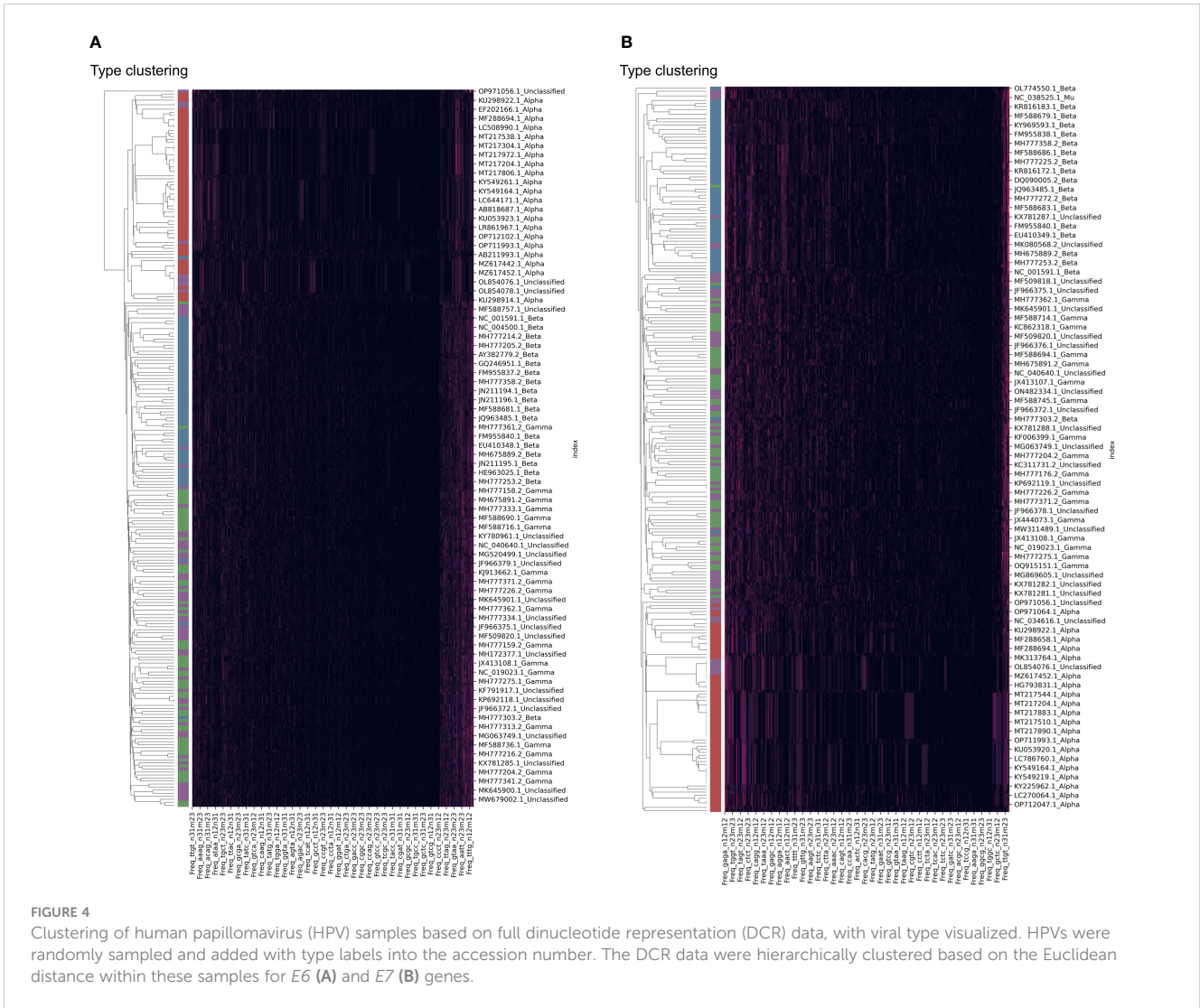
FIGURE 1

Pipeline of data cleaning and deep learning analysis of human papillomavirus (HPV) genes. **(A)** HPV data cleaning process and the sequence number for the HPVs with full *E6* and *E7* genes. **(B)** Illustration of counting method for the compositional traits of HPV *E6* and *E7* genes. **(C)** Illustration of dimension reduction and unsupervised learning of HPV genes. **(D)** The structure of convolutional neural network (CNN) model and the detailed dimension of dinucleotide representation (DCR) data during CNN. **(E)** The prediction of the oncogenicity of HPVs by the CNN classifier and the analysis of oncogenic HPVs.

The clustering of HPV samples was also performed based on the DCR trait. The HPV clustering into the main three groups of Alpha, Beta, and Gamma was interrupted by unclassified samples based on either the *E6* (Figure 4A) or *E7* (Figure 4B) gene. Therefore, it is necessary to identify the classification of these unannotated samples.

3.3 Deep learning model predicted the oncogenic potential of HPVs based on viral genomic composition of *E6* or *E7*

Given the dominant roles of *E6* and *E7* on the oncogenicity of HPVs, a deep learning predictor of the CNN model was built for the prediction of unclassified HPVs. The network structure is illustrated in Figure 1D. The DCR vector of 1,536 dimensions was sequentially

convoluted and pooled three times, and then the outputted full-connection vector of 768 dimensions was linearly transformed into a vector of 192 dimensions, with final two oncogenicity labels of 1 (oncogenicity positive) and 0 (oncogenicity negative). The DCR-based CNN classification for either *E6* or *E7* was trained. The predictor for *E6* was highly accurate for the independent HPV test data based on the ROC_AUC curve (Figure 5A); after a training epoch of 50, the accuracy was indicated by the confusion matrix (Figure 5B). The full-connection data of *E6* DCR were clearly separated based on the distribution of the two main PCA-reduced components. The PCA1 peaks of high and low oncogenic HPVs were significantly different (p < 0.0001, Figure 5C). The performance of the *E7* CNN predictor was also highly accurate. The ROC_AUC curve was almost at a right angle (Figure 5D), and the accuracy was more than 95% for low oncogenic HPVs and almost 100% for high

FIGURE 2

Counting of annotation items for human papillomaviruses (HPVs) based on original recorded information. The counting of HPVs based on labels "year" **(A)**, "country" **(B)**, "genotype" **(C)**, and "host" **(D)** was performed based on the above-mentioned annotation for each HPV record on National Center for Biotechnology Information (NCBI) website.

FIGURE 3

Plot of dinucleotide representation (DCR) data of the human papillomaviruses (HPVs), with type label or oncogenicity label. The DCR data with 1,536 dimensions were first reduced with t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA) and then were plotted with the two main reduced components, with type or oncogenicity labeled for each sample plot. **(A–D)** t-SNE- and PCA-reduced DCR for *E6* **(A, B)** or *E7* gene **(C, D)** of HPVs, with "type" labeled. **(E–H)** t-SNE- and PCA-reduced DCR for *E6* **(E, F)** or *E7* gene **(G, H)** of HPVs, with "oncogenicity" labeled.

FIGURE 4
Clustering of human papillomavirus (HPV) samples based on full dinucleotide representation (DCR) data, with viral type visualized. HPVs were randomly sampled and added with type labels into the accession number. The DCR data were hierarchically clustered based on the Euclidean distance within these samples for *E6* **(A)** and *E7* **(B)** genes.

oncogenic HPVs (Figure 5E). Similarly, the difference between the two PCA1 peaks was very significant (p < 0.0001, Figure 5F). Taken together, both *E6* and *E7* genes were greatly but vaguely different and learnable with the deep learning method.
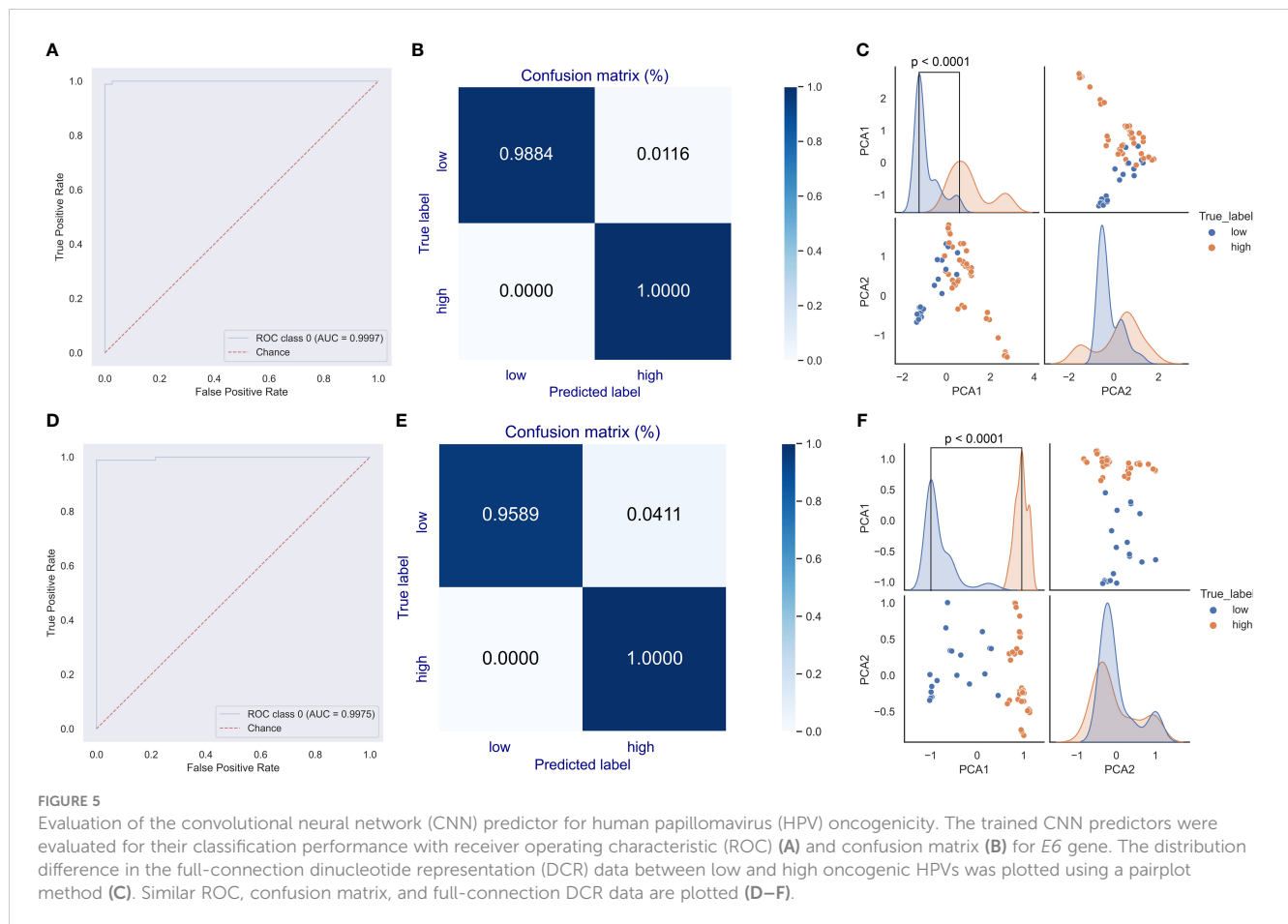
3.4 Deep learning prediction of unclassified HPVs of the oncogenic potential based on *E6* or *E7*

The two CNN predictors based on DCR data of *E6* and *E7* were utilized to predict the oncogenicity of HPVs. The 67 unclassified HPVs with full CDS of *E6* and *E7* were first counted for the DCR trait and then were input into the trained CNN predictor for either *E6* or *E7*. The predicted label of 1 or 0, the probability for each label, and the full annotation information are listed in Supplementary Table 1. As indicated, 49/67 samples were predicted to be oncogenic based on the *E6* DCR trait, whereas 9/67 unclassified HPVs were predicted to be oncogenic based on the *E7* DCR trait. The distribution of data, genotype, country, and collection site was plotted. The 49 positive

samples predicted by *E7* DCR were mainly isolated in 2018, 2017, and 2012 (Figure 6A), mostly without annotated genotype (Figure 6B), isolated in EU countries (Figure 6C), and mainly collected from skin and penile swab samples (Figure 6D). Taken together, the present deep learning predictors facilitated the oncogenicity evaluation based on viral *E6* and *E7* genes.

4 Discussion

In response to the need to evaluate the oncogenic phenotype of HPVs, particularly for the novel HPV strains, without identified genotypes, or the HPV with its oncogenicity unclear, we utilized the reported method to parse the genomic compositional traits of HPV CDSs, and trained deep learning classifiers, based on HPV *E6* or *E7* DCR data. For most types of viruses, the genome sequences do not have the same length and need to be aligned for sequence similarity analysis using Multiple Sequence Alignment (MSA). However, the high computational consumption and complexity were not balanced with an intelligent evaluation of aligned sequences. The

FIGURE 5
Evaluation of the convolutional neural network (CNN) predictor for human papillomavirus (HPV) oncogenicity. The trained CNN predictors were evaluated for their classification performance with receiver operating characteristic (ROC) **(A)** and confusion matrix **(B)** for *E6* gene. The distribution difference in the full-connection dinucleotide representation (DCR) data between low and high oncogenic HPVs was plotted using a pairplot method **(C)**. Similar ROC, confusion matrix, and full-connection DCR data are plotted **(D–F)**.

compositional DNT, DCR, and other codon-dependent traits of CDSs (Li et al., 2022, 2020) were MSA-independent and have fixed dimensions, which make further analysis easy. More importantly, such compositional traits were more biologically interpretable for virus prediction (Jiang et al., 2023; Li et al., 2023, 2022). Thus, the compositional trait of DCR with fixed dimension is more applicable to represent the genotype difference for viral genes with various sequence lengths, such as HPVs.

The oncogenicity of HPVs has been explored using AI approaches. Most of these explorations focused on diagnostic image recognition (Fu et al., 2022; La Greca et al., 2022; Klein et al., 2023; de Sanjose et al., 2024). Rare reports implicated the oncogenicity assessment of HPVs based on their genotype polymorphism. However, more and more HPV genome samples have been isolated and sequenced, and there are more novel HPV genotypes that have been identified in recent years. However, the biological identification and clinical observation of the oncogenicity of these HPV strains were much more lagged. Interestingly, the trained predictor in this study is capable of predicting the oncogenicity of these unknown genotypes, such as HPV211–215. Thus, the compositional trait-based deep learning classifiers have provided alternative and intelligent tools for HPV evaluation.

The present study analyzed the compositional characteristics of HPV *E6* and *E7* genes, which are most important for the virus oncogenicity for the first time. Our results have clearly illustrated a significant type- or genotype-dependent clustering or separation of HPVs, particularly for the high oncogenic and benign HPVs. Such marked difference in DCR distribution between the two groups was easily learned by a CNN model for the classification of the two virus groups. The high accuracy of our trained predictor on independent test data implied that the two models were reliable for the oncogenicity evaluation of HPVs. Interestingly, a much more significant difference in the DCR trait between high and low oncogenic HPVs was indicated by the PCA-reduced DCR data, implying more significance on the viral oncogenicity. The supervised learning of the CNN classifier predicted many more HPVs without clear oncogenicity, and genotyping was predicted to be oncogenic based on *E7* CDS than based on *E6*. The dominant promotion by HPV *E6* and *E7* genes to the malignant transformation of viral targeted cells has been biologically explored and widely clinically studied for a long time. However, the comparable importance between *E6* and *E7* genes is not clear. The significant difference in the oncogenicity prediction by *E6*- and *E7*-based classifiers implied different importance of the two genes.

Additionally, other HPV genes like *E2*, *E1*, *L2*, and *L1* have been indicated to be less associated with the oncogenicity of HPVs. However, their biological roles in the HPV–host interaction have less been focused, compared to *E6* and *E7*. The present study might also be beneficial for the analysis of the phenotype–genotype relationship for these genes. However, such analysis of prediction
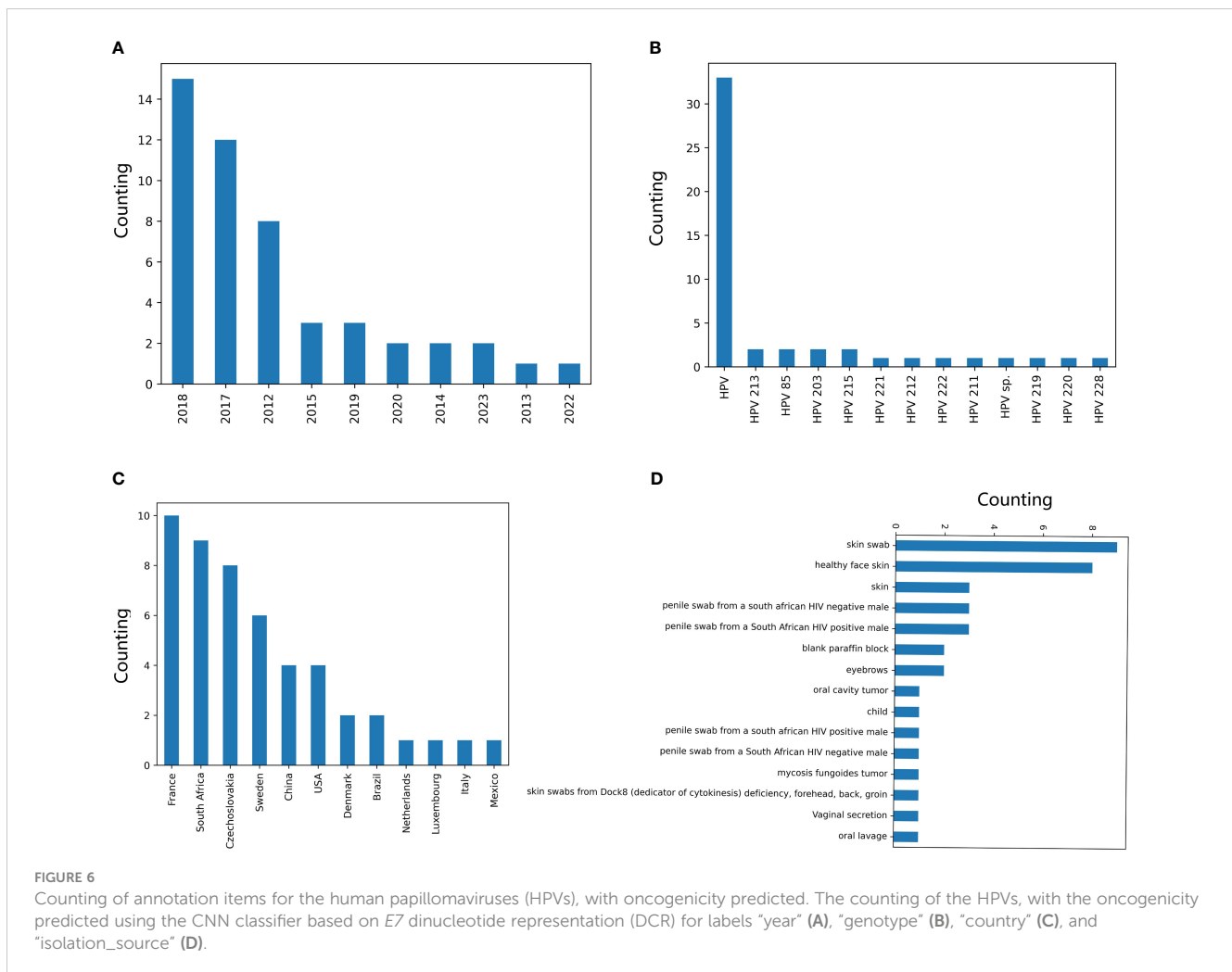
FIGURE 6
Counting of annotation items for the human papillomaviruses (HPVs), with oncogenicity predicted. The counting of the HPVs, with the oncogenicity predicted using the CNN classifier based on *E7* dinucleotide representation (DCR) for labels "year" **(A)**, "genotype" **(B)**, "country" **(C)**, and "isolation_source" **(D)**.

for the topic of oncogenicity of HPVs should be more feasible owing to the knowledge of the association between viral oncogenicity and the genotypes in these genes.

5 Conclusion

In summary, the viral genomic compositional traits characterize the oncogenicity of HPVs and facilitate the prediction of the high or low oncogenic HPVs via a DCR-based deep learning classifier for the oncogenicity-related genes *E6* and *E7*. The trained predictor in this study will facilitate the identification of HPV oncogenicity, particularly for those HPVs without clear genotype or phenotype.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

LH: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review & editing, Data curation, Visualization, Funding acquisition, Investigation, Validation. YJ: Methodology, Software, Writing – original draft, Investigation, Resources. CZ: Investigation, Methodology, Software, Writing – original draft, Visualization. PH: Investigation, Methodology, Software, Visualization, Writing – original draft, Conceptualization, Formal analysis, Project administration, Resources, Supervision, Writing – review & editing, Data curation, Funding acquisition, Validation.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2024.1430424/full#supplementary-material

References

Accardi, R., Rubino, R., Scalise, M., Gheit, T., Shahzad, N., Thomas, M., et al. (2011). E6 and E7 from human papillomavirus type 16 cooperate to target the PDZ protein Na/H exchange regulatory factor 1. *J. Virol.* 85, 8208–8216. doi: 10.1128/JVI.00114-11

Babayan, S. A., Orton, R. J., and Streicker, D. G. (2018). Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science.* 362, 577–580. doi: 10.1126/science.aap9072

Barbosa, M. S. (1996). The oncogenic role of human papillomavirus proteins. *Crit. Rev. Oncog.* 7, 1–18. doi: 10.1615/CritRevOncog.v7.i1-2

Bruyere, D., Roncarati, P., Lebeau, A., Lerho, T., Poulain, F., Hendrick, E., et al. (2023). Human papillomavirus E6/E7 oncoproteins promote radiotherapy-mediated tumor suppression by globally hijacking host DNA damage repair. *Theranostics.* 13, 1130–1149. doi: 10.7150/thno.78091

Bzhalava, D., Guan, P., Franceschi, S., Dillner, J., and Clifford, G. (2013). A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. *Virology.* 445, 224–231. doi: 10.1016/j.virol.2013.07.015

Correa, R. M., Vladimirsky, S., Heideman, D. A., Coringrato, M., Abeldano, A., Olivares, L., et al. (2017). Cutaneous human papillomavirus genotypes in different kinds of skin lesions in Argentina. *J. Med. Virol.* 89, 352–357. doi: 10.1002/jmv.24631

de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G. M. (2020). Global burden of cancer attributable to infections in 2018: A worldwide incidence analysis. *Lancet Glob Health* 8, e180–e190. doi: 10.1016/S2214-109X(19)30488-7

de Martel, C., Plummer, M., Vignat, J., and Franceschi, S. (2017). Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int. J. Cancer.* 141, 664–670. doi: 10.1002/ijc.30716

Demers, G. W., Espling, E., Harry, J. B., Etscheid, B. G., and Galloway, D. A. (1996). Abrogation of growth arrest signals by human papillomavirus type 16 E7 is mediated by sequences required for transformation. *J. Virol.* 70, 6862–6869. doi: 10.1128/jvi.70.10.6862-6869.1996

Deng, Y., Li, C., Han, J., Wen, Y., Wang, J., Hong, W., et al. (2017). Phylogenetic and genetic characterization of a 2017 clinical isolate of H7N9 virus in Guangzhou, China during the fifth epidemic wave. *Sci. China Life Sci.* 60, 1331–1339. doi: 10.1007/s11427-017-9152-1

de Sanjose, S., Perkins, R. B., Campos, N., Inturrisi, F., Egemen, D., Befano, B., et al. (2024). Design of the HPV-automated visual evaluation (PAVE) study: Validating a novel cervical screening strategy. *eLife.* 15 (12), RP91469. doi: 10.7554/eLife.91469

de Sanjose, S., Quint, W. G., Alemany, L., Geraets, D. T., Klaustermeier, J. E., Lloveras, B., et al. (2010). Human papillomavirus genotype attribution in invasive cervical cancer: A retrospective cross-sectional worldwide study. *Lancet Oncol.* 11, 1048–1056. doi: 10.1016/S1470-2045(10)70230-8

de Villiers, E. M. (1994). Human pathogenic papillomavirus types: An update. *Curr. Top. Microbiol. Immunol.* 186, 1–12. doi: 10.1007/978-3-642-78487-3_1

Drews, C. M., Case, S., and Vande, P. S. (2019). E6 proteins from high-risk HPV, low-risk HPV, and animal papillomaviruses activate the Wnt/beta-catenin pathway through E6AP-dependent degradation of NHERF1. *PloS Pathog.* 15, e1007575. doi: 10.1371/journal.ppat.1007575

Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Fu, L., Xia, W., Shi, W., Cao, G. X., Ruan, Y. T., Zhao, X. Y., et al. (2022). Deep learning based cervical screening by the cross-modal integration of colposcopy, cytology, and HPV test. *Int. J. Med. Inform.* 159, 104675. doi: 10.1016/j.ijmedinf.2021.104675

Galati, L., Chiantore, M. V., Marinaro, M., and Di Bonito, P. (2024). Human oncogenic viruses: Characteristics and prevention Strategies-Lessons learned from human papillomaviruses. *Viruses.* 16 (3), 416. doi: 10.3390/v16030416

Gheit, T. (2019). Mucosal and cutaneous human papillomavirus infections and cancer biology. *Front. Oncol.* 9. doi: 10.3389/fonc.2019.00355

Guo, L., Yang, F., Yin, Y., Liu, S., Li, P., Zhang, X., et al. (2018). Prevalence of human papillomavirus type-16 in head and neck cancer among the Chinese population: A Meta-Analysis. *Front. Oncol.* 8. doi: 10.3389/fonc.2018.00619

Jablonska, S., Dabrowski, J., and Jakubowicz, K. (1972). Epidermodysplasia verruciformis as a model in studies on the role of papovaviruses in oncogenesis. *Cancer Res.* 32, 583–589.

Jiang, S., Zhang, S., Kang, X., Feng, Y., Li, Y., Nie, M., et al. (2023). Risk assessment of the possible intermediate host role of pigs for coronaviruses with a deep learning predictor. *Viruses.* 15 (7), 1556. doi: 10.3390/v15071556

Klein, S., Wuerdemann, N., Demers, I., Kopp, C., Quantius, J., Charpentier, A., et al. (2023). Predicting HPV association using deep learning and regular H&E stains allows granular stratification of oropharyngeal cancer patients. *NPJ Digit Med.* 6, 152. doi: 10.1038/s41746-023-00901-z

La Greca, S. A., Bogowicz, M., Konukoglu, E., Riesterer, O., Balermpas, P., Guckenberger, M., et al. (2022). A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer. *Comput. Biol. Med.* 142, 105215. doi: 10.1016/j.compbiomed.2022.105215

Li, X., and Coffino, P. (1996). High-risk human papillomavirus E6 protein has two distinct binding sites within p53, of which only one determines degradation. *J. Virol.* 70, 4509–4516. doi: 10.1128/jvi.70.7.4509-4516.1996

Li, J., Tian, F., Zhang, S., Liu, S. S., Kang, X. P., Li, Y. D., et al. (2023). Genomic representation predicts an asymptotic host adaptation of bat coronaviruses using deep learning. *Front. Microbiol.* 14. doi: 10.3389/fmicb.2023.1157608

Li, J., Wu, Y. N., Zhang, S., Kang, X. P., and Jiang, T. (2022). Deep learning based on biologically interpretable genome representation predicts two types of human adaptation of SARS-CoV-2 variants. *Brief. Bioinform.* 23 (3), bbac036. doi: 10.1093/bib/bbac036

Li, J., Zhang, S., Li, B., Hu, Y., Kang, X. P., Wu, X. Y., et al. (2020). Machine learning methods for predicting Human-Adaptive influenza a viruses based on viral nucleotide compositions. *Mol. Biol. Evol.* 37, 1224–1236. doi: 10.1093/molbev/msz276

Liu, J., Nian, Q. G., Zhang, Y., Xu, L. J., Hu, Y., Li, J., et al. (2014). *In vitro* characterization of human adenovirus type 55 in comparison with its parental adenoviruses, types 11 and 14. *PloS One* 9, e100665. doi: 10.1371/journal.pone.0100665

Longworth, M. S., and Laimins, L. A. (2004). Pathogenesis of human papillomaviruses in differentiating epithelia. *Microbiol. Mol. Biol. Rev.* 68, 362–372. doi: 10.1128/MMBR.68.2.362-372.2004

Martinez-Zapien, D., Ruiz, F. X., Poirson, J., Mitschler, A., Ramirez, J., Forster, A., et al. (2016). Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. *Nature.* 529, 541–545. doi: 10.1038/nature16481

McLaughlin-Drubin, M. E., and Munger, K. (2009). Oncogenic activities of human papillomaviruses. *Virus Res.* 143, 195–208. doi: 10.1016/j.virusres.2009.06.008

Muhr, L., Eklund, C., and Dillner, J. (2018). Towards quality and order in human papillomavirus research. *Virology.* 519, 74–76. doi: 10.1016/j.virol.2018.04.003

Rashid, N. N., Rothan, H. A., and Yusoff, M. S. (2015). The association of mammalian DREAM complex and HPV16 E7 proteins. *Am. J. Cancer Res.* 5, 3525–3533.

Shu, S., Li, Z., Liu, L., Ying, X., Zhang, Y., Wang, T., et al. (2022). HPV16 E6-Activated OCT4 Promotes Cervical Cancer Progression by Suppressing p53 Expression via Co-Repressor NCOR1. *Front. Oncol.* 12. doi: 10.3389/fonc.2022.900856

Songock, W. K., Kim, S. M., and Bodily, J. M. (2017). The human papillomavirus E7 oncoprotein as a regulator of transcription. *Virus Res.* 231, 56–75. doi: 10.1016/j.virusres.2016.10.017

Sun, H., Sun, Y., Pu, J., Zhang, Y., Zhu, Q., Li, J., et al. (2014). Comparative virus replication and host innate responses in human cells infected with three prevalent clades (2.3.4, 2.3.2, and 7) of highly pathogenic avian influenza H5N1 viruses. *J. Virol.* 88, 725–729. doi: 10.1128/JVI.02510-13

Walboomers, J. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., et al. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* 189, 12–19. doi: 10.1002/(ISSN)1096-9896

White, E. A., Sowa, M. E., Tan, M. J., Jeudy, S., Hayes, S. D., Santha, S., et al. (2012). Systematic identification of interactions between host cell proteins and E7 oncoproteins from diverse human papillomaviruses. *Proc. Natl. Acad. Sci. U S A.* 109, E260–E267. doi: 10.1073/pnas.1116776109

Whitley, R. J., and Roizman, B. (2001). Herpes simplex virus infections. *Lancet.* 357, 1513–1518. doi: 10.1016/S0140-6736(00)04638-9

Yanatatsaneejit, P., Chalertpet, K., Sukbhattee, J., Nuchcharoen, I., Phumcharoen, P., and Mutirangura, A. (2020). Promoter methylation of tumor suppressor genes induced by human papillomavirus in cervical cancer. *Oncol. Lett.* 20, 955–961. doi: 10.3892/ol.2020.11625

Zhang, B., Chen, W., and Roman, A. (2006). The E7 proteins of low- and high-risk human papillomaviruses share the ability to target the pRB family member p130 for degradation. *Proc. Natl. Acad. Sci. U S A.* 103, 437–442. doi: 10.1073/pnas.0510012103

Zur, H. H. (2000). Papillomaviruses causing cancer: Evasion from host-cell control in early events in carcinogenesis. *J. Natl. Cancer Inst.* 92, 690–698. doi: 10.1093/jnci/92.9.690

Zur, H. H. (2002). Papillomaviruses and cancer: From basic studies to clinical application. *Nat. Rev. Cancer.* 2, 342–350. doi: 10.1038/nrc798