Genetic characteristics of human papillomavirus type 16, 18, 52 and 58 in southern China

Running Title: Genetics of four HPVs in southern China

Yuee Zu,[1#] Zhihua Ou,[2,3#] Dan Wu,[1#] Wei Liu,[2,3,4#] Liwen Liu,[5#] Di Wu,[2,3,6#] Yanping Zhao,[2,3#] Peidi Ren,[2,3#] Yanqing Zhang,[1] Wangsheng Li,[2,7,8] Shujin Fu,[2] Yongchun Wen,[1] Xianchu Cai,[2,7,8] Wenbo Liao,[1] Chunyu Geng,[2] Hongcheng Zhou, [2,7,8] Xiaman Wang,[2,9] Haorong Lu,[2,7,8] Huanhuan Peng,[2,9] Na Liu,[2,9] Shida Zhu,[2,9,10] Jiyang Liu,[5*] Dongbo Wang,[1*] Junhua Li[2,3,4*]

[1] Changsha Maternal and Child Care Hospital, Changsha 410003, Hunan, China.

[2] BGI-Shenzhen, Shenzhen 518083, China.

[3] Shenzhen Key Laboratory of Unknown Pathogen Identification, BGI-Shenzhen, Shenzhen 518083, China.

[4] School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China.

[5] The First Hospital of Changsha City, Changsha, Hunan 410005, China.

[6] School of Basic Medicine, Qingdao University, Qingdao 266071, China.

[7] China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China.

[8] Shenzhen Key Laboratory of Environmental Microbial Genomics and Application, BGI-Shenzhen, Shenzhen 518083, China.

[9] BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China.

[10] Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, BGI-Shenzhen, Shenzhen 518120, China.

[#]These authors contributed equally to this work. *Corresponding author.

Word counts of the abstract: 148

Word counts of the text: 4957

Abstract

Persistent infections of high-risk human papillomaviruses (HPVs) are the leading cause of cervical cancers. We collected cervical exfoliated cell samples from females in Changsha city, Hunan Province and obtained 358 viral genomes of four major HPV types, including HPV 16 (n=82), 18 (n=35), 52 (n=121) and 58 (n=100). The lineage/sublineage distribution of the four HPVs confirmed previous epidemiological reports, with the predominant prevailing sublineage as A4 (50%), A1 (37%) and A3 (13%) for HPV16, A1 (83%) for HPV18, B2 (86%) for HPV52 and A1 (65%), A3 (19%) and A2 (12%) for HPV58. We also identified two potentially novel HPV18 sublineages, i.e. A6 and A7. Virus mutation analysis further revealed the presence of HPV16 and HPV58 strains associated with potentially high oncogenicity. These findings expanded our knowledge on the HPV genetic diversity in China, providing valuable evidence to facilitate HPV DNA screening, vaccine effectiveness evaluation and control strategy development.

Keywords: HPV; high risk; Chinese female; cervical cancer; lineage; sublineage; complete genome.

Introduction

Human papillomaviruses (HPVs) are double-stranded DNA viruses responsible for nearly all cervical cancers and related to multiple types of other cancers [1]. Of a big family with over 200 different genotypes identified so far, more than 12 HPV types were carcinogenic, among which HPV16 and HPV18 were responsible for over 70% of the cervical cancer cases [2–4]. Meanwhile, HPV52 and HPV58 were dominantly prevailing in East Asian countries, causing more precancer and invasive cancer in this region than elsewhere [5]. Cervical cancer, as the fourth major cancer in women, affected an estimated 570,000 individuals and caused 311,000 deaths in 2018 [6]. To eliminate cervical cancer as a global public health problem, the World Health Organization has raised several targets to be fulfilled by every country by 2035 [7], including 90% vaccination coverage among girls below 15 years of age, 70% screening coverage in 35-45-year-old women and 90% treatment coverage for precancerous and cancer patients. However, the implementation between high-income and low-income countries is significantly different. While 85% of those girls in high-income countries have been covered by national HPV vaccination programs already, only 20-30% of those in low or medium-income countries are covered. Screening implementation also encounters the same dilemma. Therefore, 90% of the cervical cancer deaths occurred in low- or mid-income countries where the preventive measures are least widely practiced [8,9]. A sharp contrast could be found between the US and China. In 2015, the incidence of cervical cancer was 12,900 with 4,100 deaths in the US, while in China, the numbers were estimated as 98,900 and 30,500 respectively [10,11]. The high incidence and mortality of cervical cancer in China was due to the low coverage in both HPV vaccination and cervical cancer screening, under which coverage, the incidence of cervical cancer in 2100 was projected as three times that of 2015 [12]. To meet the WHO target of cervical cancer elimination, both China and other low- and medium-income countries have a long way to go.

The HPV genomes usually encode one long control region (LCR), six open reading frames (ORFs) containing E1, E2, E4, E5, E6 and E7, and the late ORFs expressing L1 and L2 capsid proteins [13]. To better distinguish viral heterogeneity, researchers have designated HPV of the same type into lineage and sublineages, which requires complete genomic nucleotide differences of 1%-10% and 0.5%-1%, respectively [14]. Four lineages (A, B, C and D) and sixteen sublineages of HPV16 (A1-A4, B1-B4, C1-C4, D1-D4) [14–16], three lineages (A, B and C) and nine sublineages (A1-A5, B1-B3, C) of HPV18 [14], four lineages and seven sublineages (A1-A2, B1- B2, C1-C2, D) of HPV52 [14], and four lineages and eight sublineages (A1-A3, B1-B2, C, D1-D2) of HPV58 have been identified so far [14].

HPV sublineages varied both in geographical regions and carcinogenicity. Based on over 7,000 HPV16 positive samples from 52 countries, it was found that the predominant global sublineage was A1, though some other regional predominant sublineages also existed [15]. In East Asia, the regional prevalent HPV16 sublineages were A3 and A4, and sublineage A3, A4 and lineage D may have higher cancer risks than A1 and A2 [15–17]. In Africa, while HPV16 A1/A2 were still prevalent, lineages B, C and D were also common [15]. HPV18 also displayed geographical heterogeneity, with lineage A dominant around the world except in Africa where lineage B was more popular. Within HPV18 lineage A, sublineage A1 was predominant in East Asian and the Pacific [18]. No distinctive cancer risks were identified among the different lineage/sublineages of HPV18, probably due to the limited sample sizes involved in the investigations [18]. As for HPV52, lineage B was predominant in Asia, while lineage A was the predominant lineage in the other continents [19]. For HPV58, lineage A (especially A2) was predominant globally, but lineage C and D were also common in Africa, accounting for 39.1% and 8.7% of all the samples from Africa [20]. It was also reported that the HPV58 A3 sublineage

104 containing E7:T20I/G63S mutations had higher oncogenicity [21,22]. While full
105 genomes were not always available for lineage/sublineage study, most of the above
106 studies used partial genomes to designate HPV lineages/sublineages.
107
108 Multiple epidemiological investigations have shown that HPV52 and HPV58 were
109 among the most common HPVs infecting woman cervix in China, and may cause
110 more infections than HPV16 and HPV18 in some regions [23–25]. The disease
111 burden caused by HPV52 and HPV58 should not be neglected considering the large
112 population size in China, although their carcinogenicity was inferior to that of HPV16
113 and HPV18 [26]. Currently, most of the investigations on HPV genomes were carried
114 out in North America, only a small amount of them were from China. Moreover, the
115 lineage/sublineage distribution of HPV types was mainly classified by partial
116 genomes or single genes, and the HPV diversity in southern China remained
117 obscure. Herein, using samples from Changsha, Hunan Province, we aimed to
118 reveal the genetic diversity of the HPV16, 18, 52 and 58 with full viral genome
119 sequences. A comprehensive understanding of the viral characteristics of these four
120 types would facilitate the evaluation of the protection potency of related HPV
121 vaccines in China.
122
123 **Results**
124 **Classification of HPV16 genomes from Changsha**
125 In this study, we have obtained 82 HPV16 full genomes from Changsha (Table 1,
126 Supplementary Table S1). Although HPV16 was the most common HPV type
127 causing cervical cancer, HPV16 genomes from China were limited. Combined with
128 public data, we constructed Maximum-Likelihood phylogeny based on full viral
129 genomes to understand the lineage distribution of the Changsha strains. Based on
130 phylogeny clustering and distance comparison, the HPV16 from Changsha city were
131 mainly assigned to sublineage A4 (41, 50.00%), A1 (30, 36.59%), and A3 (11,

13.41%) (Table 1, Figure 1). Some of the strains in sublineage A1 formed two new branches neighboring A3 and A4, which were labelled as cluster M and N in the HPV16 phylogeny (Figure 1). Pairwise sequence distance calculation showed that genomes in the two branches differed from the A1 reference genome by less than 0.5%, indicating that the two branches still belonged to sublineage A1. Besides the sequences from China, HPV16 strains from other countries including the United States, Japan and Thailand, were also identified in these two branches (Figure 1, Supplementary Figure S1).

Classification of HPV18 genomes from Changsha

For HPV18, we have obtained 35 full genomes from samples in Changsha (Table 1, Supplementary Table S1). Based on phylogeny topology and nucleotide difference, the HPV18 from Changsha city were mainly assigned to sublineage A1 (29, 82.86%). We also identified one strain each for sublineages A3, A4 and A5. Moreover, there were three strains that belonged to lineage A but not assigned to any knonw sublineages (Table 1, Supplementary Table S1, Supplementary Figure S2), which we designated as sublineages A6 and A7 (Figure 2). Sublineage A6 contained three strains, with two from China and one from the Netherlands. These strains had less than 0.5% distance from HPV18 sublineage A3 reference genome but failed to form a monophyletic branch with sublineage A3. Meanwhile, their nucleotide differences from sublineages A1, A2 and A4 were approximately 0.5%. Therefore, we decided to classify this cluster as a new sublineage. Sublineage A7 was a neighboring branch to HPV18 sublineage A5, which contained two strains from Changsha with sequence distances of 0.5%-1% from all the A sublineages. Considering only limited strains were present in sublineage A6 and A7, more epidemiological evidence may be needed to support the presence of the two new HPV18 sublineages.

Classification of HPV52 genomes from Changsha

For HPV52, we have obtained 121 full genomes from samples in Changsha (Table 1, Supplementary Table S1). The HPV52 strains were mainly assigned to sublineage B2 (104, 85.96%, Figure 3, Supplementary Figure S3). Several strains from sublineage A1 (5, 4.13%), C2 (4, 3.31%) and D1 (8, 6.61%) were also detected, implicating the sporadic circulation of multiple lineages in this region. It was intriguing to have HPV52 sublineage D1 detected, which was rarely reported by other studies in Asia.

Classification of HPV58 genomes from Changsha

For HPV58, we obtained 100 full genomes from the Changsha samples (Table 1, Supplementary Table S1). The HPV58 from Changsha city were mainly assigned to sublineage A1 (65, 65.00%), A2 (12, 12.00%) and A3 (19, 19.00%) (Figure 4, Supplementary Figure S4). Four strains belonged to sublineage B1 were also detected.

Lineage/sublineage conservative mutations identified for HPV16, 18, 52 and 58

Lots of the epidemiological studies on the lineage or sublineage distributions of HPVs were relied on partial genome sequencing and comparison on nucleotide polymorphism. To help refine fixed mutations of lineages or sublineages, we combined the Changsha data with publicly available HPV genomes to conduct genome-wide mutation analysis for the four high-risk HPV types (Supplementary Table S2). Herein, we defined the mutations that occurred in over 98% of the strains belonging to at least one lineage or sublineage as conservative mutations. Mutations in E4 were not shown in the main text because this gene located within E2, but the related details could be found in the supplementary materials.

For HPV16, a total of 210 positions with conservative mutations were identified based on 2,480 full genomes (Table 2, Supplementary Table S3), these included 79 missense and 91 synonymous mutations occurred in seven genes (E1, E2, E5, E6, E7, L1 and L2), and 40 point mutations occurred in the long control region (LCR) and noncoding regions (NCR). Among them, 39 mutations were probably unique to 11 sublineages (Table 3). No unique mutations were identified for sublineages A2, A3 and B1. L1, L2 and LCR contained the most abundant unique mutations, and could be used to distinguish 7, 6 and 6 HPV16 sublineages, respectively. The combination of the unique sites from L1, L2 and LCR would be able to distinguish at least 11 sublineages.

For HPV18, a total of 269 positions with conservative mutations occurring in at least one lineage or sublineage were identified based on 182 full genomes (Table 2, Supplementary Table S4). These mutations included 76 missense mutations, 125 synonymous mutations and 6 deletions in the coding genes, 37 point mutations and 25 deleted sites occurred in the NCR and LCR (The deletion at one nucleotide position was counted as one deletion site. For example, a 6bp consecutive deletion was counted as 6 deletion sites). A 6bp deletion in the overlapping region of E2, CCTACA3630-3635del (E2:PT272-273del), and A 7bp deletion in LCR (TGTTGTA7234-7240del) was observed in sublineages A5, A7, B1, B2 and B3 and lineage C (Supplementary Table S4). In NCR, an 18bp deletion (3915-3932) occurred in strains of sublineages B1 and B2, and a 7bp deletion (3919-3925) was observed in lineage C only. A total of 74 lineage/sublineage unique mutations were identified (Table 3). L1 and L2 contained the most abundant unique positions, and their combination could distinguish all the HPV18 lineages/sublineages. For the two novel HPV18 sublineages, i.e., A6 and A7, five and ten unique mutations were identified respectively. Because some sublineages of HPV18 contained very few genomes, such as sublineages A5, A6, A7 and lineage C (Supplementary Table S2),

it's possible that some random mutations might be included. Still, for some major sublineages, the unique mutations would be informative, such as T5619A(L1:L64M) for sublineage A3, T1452C(E1:V180A) and C4341A(L2:P33H) for sublineage A4, A5924T(L1:G165G) for sublineage B1, etc.

For HPV52, a total of 289 positions with conservative mutations were identified based on 314 full genomes (Table 2, Supplementary Table S5). These mutations included 66 missense mutations and 107 synonymous mutations in the coding genes, 45 point mutations and 40 deleted sites occurred in the NCR and LCR, and 31 inserted nucleotides in L1 and NCR. Interestingly, a 21bp deletion (4169-4189) in NCR and an 8bp deletion in LCR (7700-7707) was only observed in lineage D (Table 3， Supplementary Table S5). We also found a long insertion comprised of 28 nucleotides between the 4160th and 4161st positions of HPV52 (based on the A1 reference genome), which occurred in all strains of lineage B, C and D. Moreover, lineage D had G at the 21st position of this long insertion, while lineage B and C had T (Table 3). A total of 169 unique mutations were identified. E1, E2, L1 and LCR all contained unique mutations for five sublineages, and the combination of unique sites from L1 and LCR would be able to distinguish all the lineages and sublineages of HPV52. It should be noted that lots of the unique sites belonged to lineage D, which was probably due to the long genetic distances between lineage D and the other lineages.

For HPV58, a total of 186 positions with conservative mutations were identified based on 321 full genomes (Table 2, Supplementary Table S2, Supplementary Table S6). These mutations included 50 missense mutations and 81 synonymous mutations in the coding genes, 43 point mutations in NCR and LCR, and 12 deleted sites in LCR. A 7bp consecutive insertion (7164-7170) in LCR was found in sublineages D1 and D2. A total of 79 unique sites was identified for HPV58, and E1

243 contained the most abundant unique sites for all the lineages and sublineages (Table
244 3).
245
246 **Discussion**
247 Previous studies on HPV sublineages were mainly conducted using partial gene
248 sequences and may fail to reveal the variation profiles of full genomes. In this study,
249 we have conducted genomic surveillance on four high-risk HPV types in Changsha
250 city to explore the genetic diversity of HPV16, 18, 52 and 52 in southern China in a
251 much higher resolution using complete viral genomes. We showed that A4, A1 and
252 A3 were the major HPV16 sublineages circulating in Changsha, similar to our
253 previous findings based on samples from eastern China [27]. Our genomic findings
254 were also consistent with the epidemiological reports by other studies using partial
255 gene sequencing [28,29]. We have also identified two minor clusters within HPV16
256 sublineage A1. While the two clusters were mainly formed by strains from China,
257 Japan and Thailand (Supplementary Figure S1), the exact distribution of cluster M
258 and N in East Asian countries remains to be clarified. Expanding genomic
259 surveillance in Asia would further reveal the geographical structure of HPV16
260 variants. Complete genomes of HPV18, 52 and 58 were relatively limited in public
261 database, with 150-250 genomes of each type from Asia, Europe and the Americas
262 available for this investigation [27]. Our study showed high prevalence of HPV18
263 sublineage A1 in Changsha, China (Table 1), in accordance with findings by Chen et
264 al [18]. We have also detected two potential novel sublineages of HPV18 in this
265 region (Figure 2, Supplementary Figure S2), which was not identified in other
266 regions. Interestingly, our previous genomic study on HPV18 in eastern China
267 showed a Chinese cluster in sublineage A4 [27], which was not found in this study
268 (Supplementary Figure S2). This suggests that the genetic diversity of HPV18 may
269 be more divergent than what has been reported and that different regions in China
270 may have their unique variants. HPV52 displayed higher divergence than HPV16

and HPV18, as the isolates belonged to four different lineages. Based on available genome data, B2 was the most prevalent sublineage for HPV52 both in China and the world (Supplementary Figure S3) [30,31]. Although in low ratios, HPV52 strains of lineage C and D were continuously detected in China [27,32], whether this is due to long-term maintenance or recent population movement remains unknown. HPV58 mainly belonged to lineage A, with A1 as the dominant sublineage (Supplementary Figure S4), consistent with the overall distribution of HPV58 in Asia [20,30,32]. Our phylogenetic classification revealed several potentially high-risk HPV sublineages in Changsha, including HPV16 A4 [16] and HPV58 A3 [33], raising concerns about their circulation status among the Chinese population. Our study provided useful information on the sublineage distribution of four major high-risk HPV types, which would serve as useful resources for the surveillance and control of HPVs in Chinese females.

The sublineage conservative mutations identified in our study also provided evidence on the genetic divergence and potential high-risk markers in the HPV variants in China. For example, the T178G/A (E6: D25E/E) mutation, especially T178G, has been repeatedly detected in prevailing HPV16 variants in many provinces of China, including Hubei, Xinjiang, Liaoning, Harbin, Zhejiang, Yunnan, Taiwan, Hong Kong, etc [29,34–38]. Here, T178G was uniquely detected in 40 out of 41 (97.6%) sublineage A4 strains from Changsha, while T178A was uniquely detected in 7 out of 11 (63.6%) strains of sublineage A3, suggesting that T178G mutation may be a common mutation for sublineage A4. Therefore, it is possible that the variants containing T178G mutations belong to HPV16 sublineage A4 and that this sublineage may be prevalent in different provinces of China. HPV16 T350G (E6:L83V) was another point mutation that's frequently mentioned in HPV16 variants [17,37,39], which might be a marker for sublineage D3 [35]. While no genomes of sublineage D3 were obtained from Changsha, we did find 3 strains (10%) of

sublineage A1 with the T350G mutation. Moreover, this mutation was also found in sublineage A2 and B4 (Supplementary Table S3), indicating that T350G was not specific to assign lineages or sublineages. It was suggested that HPV16 variants with the E6:D25E mutation in the Japanese population and those with E6:L83V mutation in Swedish women were linked with higher carcinogenicity, but the sample sizes of such studies were relatively small [40,41]. For HPV18, mutations related to carcinogenicity were rarely reported. For HPV52, Choi *et al* reported that A379G (E6 K93R) might be linked with higher disease severity and that this mutation was mainly occurred in sublineage B2 [31]. In our study, we showed that A379G mutation occurred in not only sublineage B2, but also in A2 and B1 (Supplementary Table S5). Other mutations previously found to be associated with sublineage C, such as A706G, G707A, T727G, C733T, G742A and T848G [31], were also detected in the HPV52 strains (n=4) of sublineage C2 in Changsha. Moreover, these mutations were also found to be unique to lineage C (Supplementary Table S5). For HPV58, both epidemiological and experimental evidence indicated that linked mutations of C632T (E7:T20I) and G760A (E7:G63S), mostly found in sublineage A3, were positively associated with high oncogenicity [22,33,42]. While C632T uniquely occurred in 100% (n=19) of the Changsha sublineage A3 strains, G760A was found in all strains of sublineage A3 (n=19) and B1 (n=4) (Supplementary Table S6). Unfortunately, our dataset failed to provide supportive evidence regarding the carcinogenicity of HPV sublineages. Zehbe *et al* pointed out that the carcinogenicity of HPV variants may also depend on the host immune system, for example, the individual difference in the major histocompatibility complex [41]. Current study also showed that some of the risk-associated mutations, such as HPV16 E6:L83V, were not lineage- or sublineage-specific. Whether the lineage/sublineage classification or mutation detection better relates with disease progression remains to be explored. Moreover, with the availability of more abundant complete viral genomes, the lineage- or sublineage-specific mutations would also be further refined.

With an aim to understand the genetic diversity of the four major high-risk HPVs, the samples were randomly selected based on their infection types but not strictly taken in accordance with their epidemiological prevalence. Although the infection and clinical information for samples were summarized (Table 1), such data were not eligible for any correlation analysis, which presents as one limitation of this study. In our mutation analysis, the lineages or sublineages containing the fewer stains tended to have more unique mutations. Certain unique mutations identified for the minor lineages/sublineages may be false positive but could be refined by increasing the dataset in the future. The mutation profile might help reveal the genetic divergence of the HPV strains in southern China but some of them might not be common characteristics for strains from other areas.

Due to the low mutation rate of DNA viruses and the transmission route of HPVs, our study in Changsha would help reveal the HPV genomic heterogeneity in Hunan and even southern China. However, for regions with unique ethnic groups and coastal regions with frequent population exchange, the HPV diversity may be more complicated. Moreover, HPVs displayed the highest diversity in Africa [15,20,43], where genomic surveillance was relatively limited. With the initiation of the cervical cancer elimination campaign, continuing HPV screening and gradual application of vaccines are expected around the world. It would be critical to understand the viral genetics before and after vaccine coverage, so as to better evaluate the effectiveness of vaccines, make appropriate adjustments to vaccine combination of HPV types, select the best reference vaccine candidates and improve other related preventive strategies. The HPV genomes generated by our study would serve as valuable baseline reference data for the control of HPVs in women in southern China. The conservative mutations identified may also facilitate large-scale lineage/sublineage classification of HPV16, 18 52 and 58 during epidemiological

study. Conducting genomic surveillance in different regions of China and other regions of the world would help us better understand the baseline of viral activity, hence deciding the best practice for screening strategies and vaccine application.

Methods and Materials

Sample collection

Exfoliated cervical cell samples were obtained from women aged from 35 to 64 participating in the Cervical Cancer Screening Program from May to June 2019 in Changsha City, Hunan Province, China. HPV types were determined with BGI SeqHPV Kit (BGI-Shenzhen, China). All the participants consented to the donation of their leftover clinical samples for this investigation. The samples that were positive of HPV16, 18, 52 or 58, regardless of single-type or multiple-type infections, were used for DNA extraction.

HPV DNA enrichment and sequencing

Genomic DNA was extracted with MagPure Buffy Coat DNA Midi KF Kit (Magen, China). Samples with a total DNA amount of over 400ng were further fragmented to around 200–400 bp with a Covaris LE-220 ultrasonicator. These fragments were then used to construct the sequencing libraries following the instruction of the MGIEasy DNA Library Preparation Reagent Kit (MGI, BGI-Shenzhen, China). These purified fragments were end repaired, adaptor-ligated, and amplificated with 8 PCR cycles. The post-PCR products were used to enrich HPV fragments. HPV RNA probes targeting 18 HPV types (HPV6, 11, 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 69 and 82) were designed by MyGenostics [44] and synthesized by MGI, BGI-Shenzhen. The hybridization process was carried out according to the manufacturer's instructions (MGI, Shenzhen, China). Libraries were hybridized with HPV probes at 65 °C for 24 hours to capture HPV-specific fragments. The eluted

fragments were amplified by 18 cycles of PCR and then used to generate DNA nanoball-based libraries after circularization and rolling circle amplification. Libraries were sequenced as 100bp paired-end reads on the BGISEQ-500 and MGISEQ-T1 sequencing platform (MGI, Shenzhen, China).

Complete genome assembly

The raw reads were quality-checked and trimmed with fastp [45]. Deduplicated reads were further removed with BBMap (https://sourceforge.net/projects/bbmap/). The clean reads were mapped to HPV reference genomes with the BWA alignment tool [46]. Reads with both ends aligned to HPV reference genomes were extracted and used for *de novo* assembly with NOVOPlasty [47]. The assembled contigs were locally blasted against a reference genome of the same type to identify the relative genomic location. The genomic fragments were then adjusted to have the same genomic coordination as the reference genome. Reference genomes for the four HPV types investigated in this study were: HPV16, K02718; HPV18, AY262282; HPV52, X74481; HPV58, D90400. It should be noted that the HPV16 reference strain K02718 was downloaded from PaVE [4], which was longer than the GenBank version by two nucleotides.

Phylogeny reconstruction and data visualization

The best-fit nucleotide substitution models were determined by ModelFinder and the Maximum Likelihood (ML) phylogenies were all constructed with 1,000 implementations of ultrafast bootstrap tests with IQ-TREE . To fully reveal the genetic diversity of the new viral genomes generated by this study, we initially constructed maximum phylogenies of the four HPV types combining genomes from NCBI and CNSA for our preliminary analysis. The total sequences and the nucleotide substitution models were: HPV16, n=2584, GTR+F+I+G4; HPV18, n=182, TVM+F+I+G4; HPV52, n=315, GTR+F+I+G4; HPV58, n=323, TVM+F+I+G4. To

improve visual clarity, we selected representative strains to construct the final phylogenies: HPV16, n=83, TVM+F+I+G4; HPV18, n=65, K3Pu+F+I; HPV52, n=66, TVM+F+I; HPV58, n=55, TVM+F+I. The pairwise nucleotide differences were calculated with *seqinR (http://seqinr.r-forge.r-project.org/)*. Based on genomic differences and phylogenetic topology, lineage and sublineage classification were performed for four HPV types [14]. Visualization of phylogeny and the associated data were carried out with *ggtree* package in R [48].

Mutation detection

Mutation detection was conducted by comparing the new genome sequences against the A1 reference genome of each HPV type (HPV16, K02718 from PaVE [4], with E6 started from genomic nucleotide position 104, which was different from the GenBank record; HPV18, AY262282; HPV52, X74481; HPV58, D90400). Only genomes with correct reading frames for all the eight coding regions (E1, E2, E4, E5, E6, E7, L1 and L2) were used for mutation detection. To remove random mutations, only those detected in over 98% sequences of any lineage or sublineage were retained. Mutations that uniquely detected in a specific lineage or sublineage were further identified. Sequences with incorrect reading frame or early stop codons in the coding regions were excluded from the mutation analysis. The numbers of genomes used for mutation analysis were as follows: HPV16, n=2480; HPV18, n=182; HPV52, n=314; HPV58, n=321 (Supplementary Table S2).

Ethical statement

This study was reviewed and approved by the Institutional Review Board of both Changsha Maternal and Child Care Hospital and Beijing Genomics Institute, Shenzhen, China (BGI-R071-1-T1 & BGI-R071-1-T2). All the participants provided written consent for this study.

Data availability

The data that support the findings of this study have been deposited into CNSA (CNGB Sequence Archive) of CNGBdb under project number CNP0001700 (https://db.cngb.org/cnsa/).

Author contributions

Junhua Li, Yuee Zu, Zhihua Ou, Jiyang Liu and Dongbo Wu designed and supervised the study. Yuee Zu, Dan Wu, Liwen Liu, Wenbo Liao, Yanqing Zhang, Yongchun Wen, Zhihua Ou, Yanping Zhao, Xiaman Wang, Huanhuan Peng, Na Liu and Shida Zhu coordinated sample collection. Peidi Ren, Wei Liu and Hongchen Zhou performed DNA extraction and library construction. Wangsheng Li, Shujin Fu, Chunyu Geng, Xianchu Cai and Haorong Lu conducted viral genome sequencing. Zhihua Ou, Di Wu and Wei Liu conducted data analysis. Zhihua Ou, Yanping Zhao, Peidi Ren and Wei Liu wrote the manuscript. Junhua Li, Yuee Zu, Dan Wu and Liwen Liu provided critical comments on the manuscript.

Declarations of interest

The authors declare no conflict of interest.

Funding

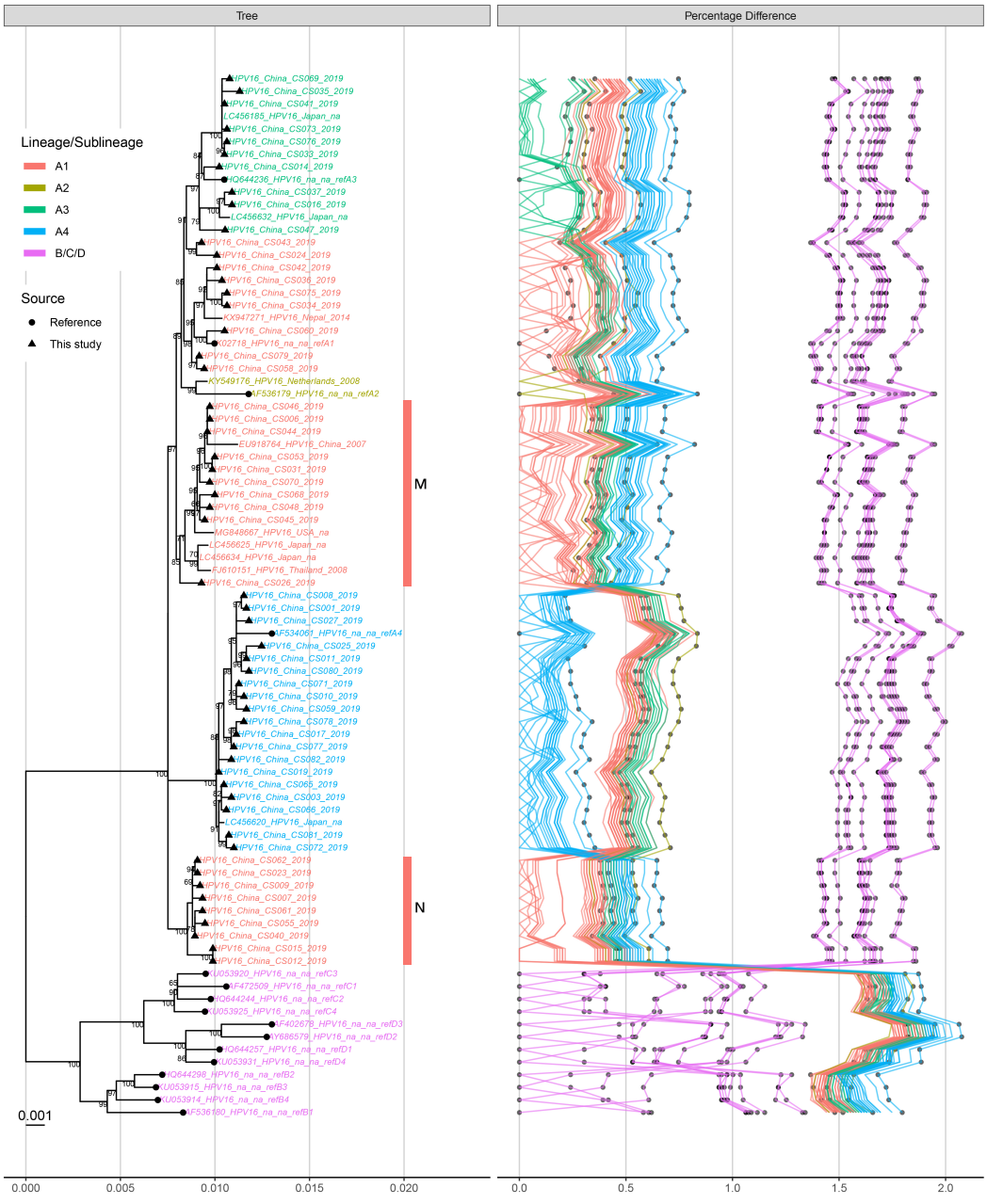
This work received no funding.

Acknowledgments

We thank China National GeneBank for providing sequencing service for this project. We warmly thank Ms. Jieyao Yu, Ms. Wei Zhou and Mr. Qineng Li for their assistance in viral genome sequencing. The authors also wish to extend their gratitude to Miss Feiyun Ou and Mr. Geer Xi for their inspirational communications.

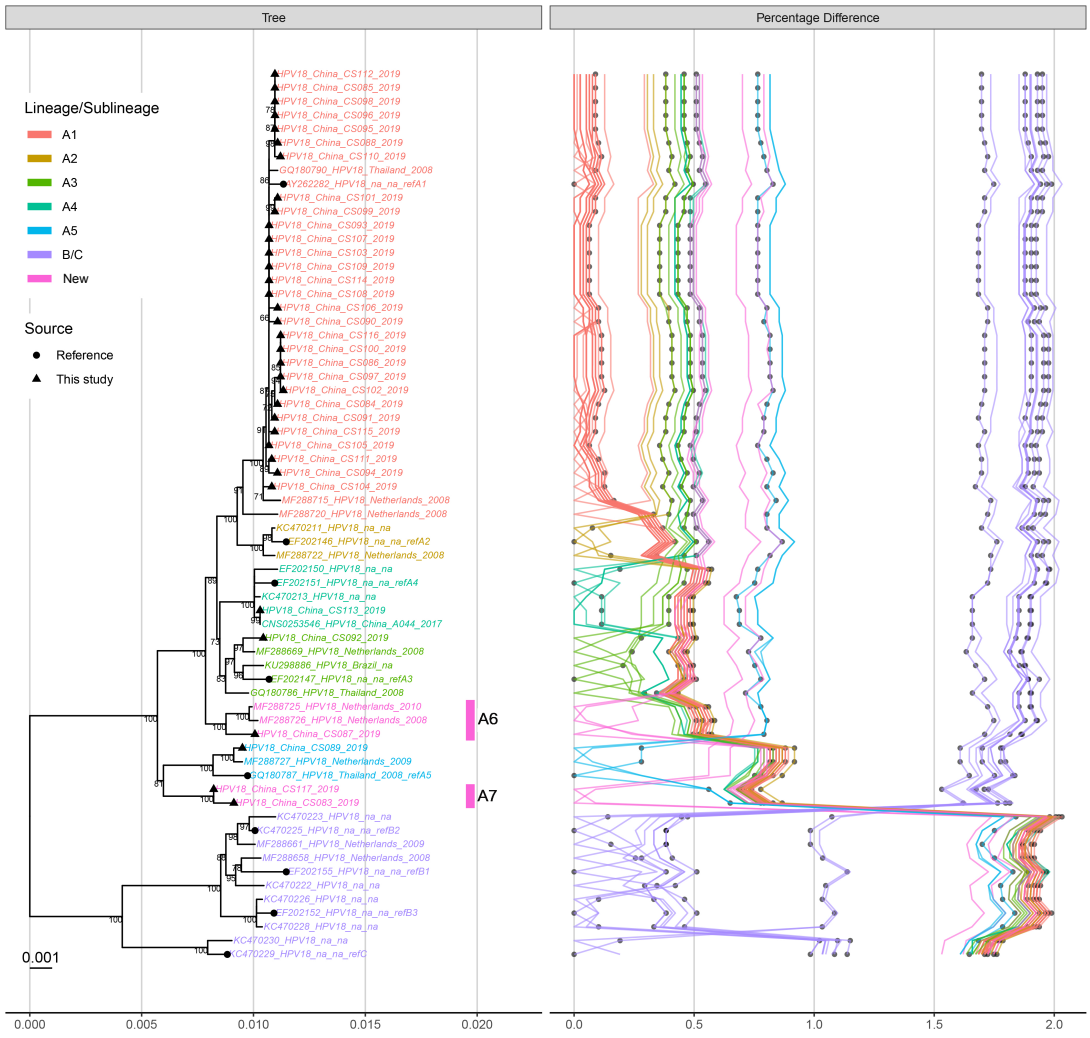
Figures and Tables

Figure 1：Representative Maximum Likelihood phylogeny of HPV16. HPV genomes generated by this study were combined with those from public database to construct a Maximum Likelihood phylogeny with 1,000 bootstrap tests. Both the phylogenetic tree (left panel) and the pairwise sequence distance (right panel) are shown. Sources of the sequences are indicated by shapes while lineages or sublineages are distinguished by different colors. Bootstrap values over 70 are labeled on nodes.



476 **Figure 2：Representative Maximum Likelihood phylogeny of HPV18.** Figure

477 legends are the same as Figure 1.

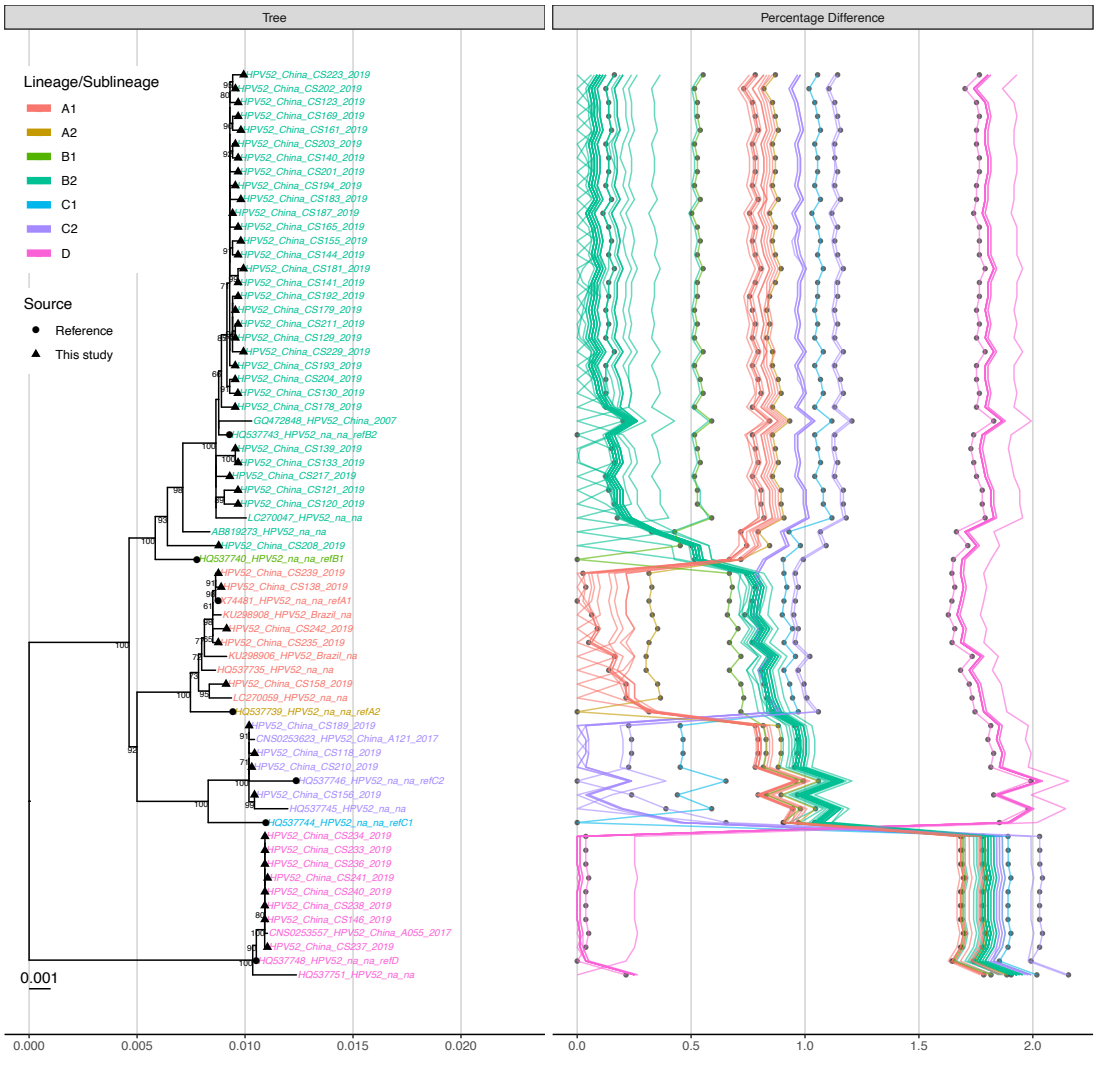


478

479

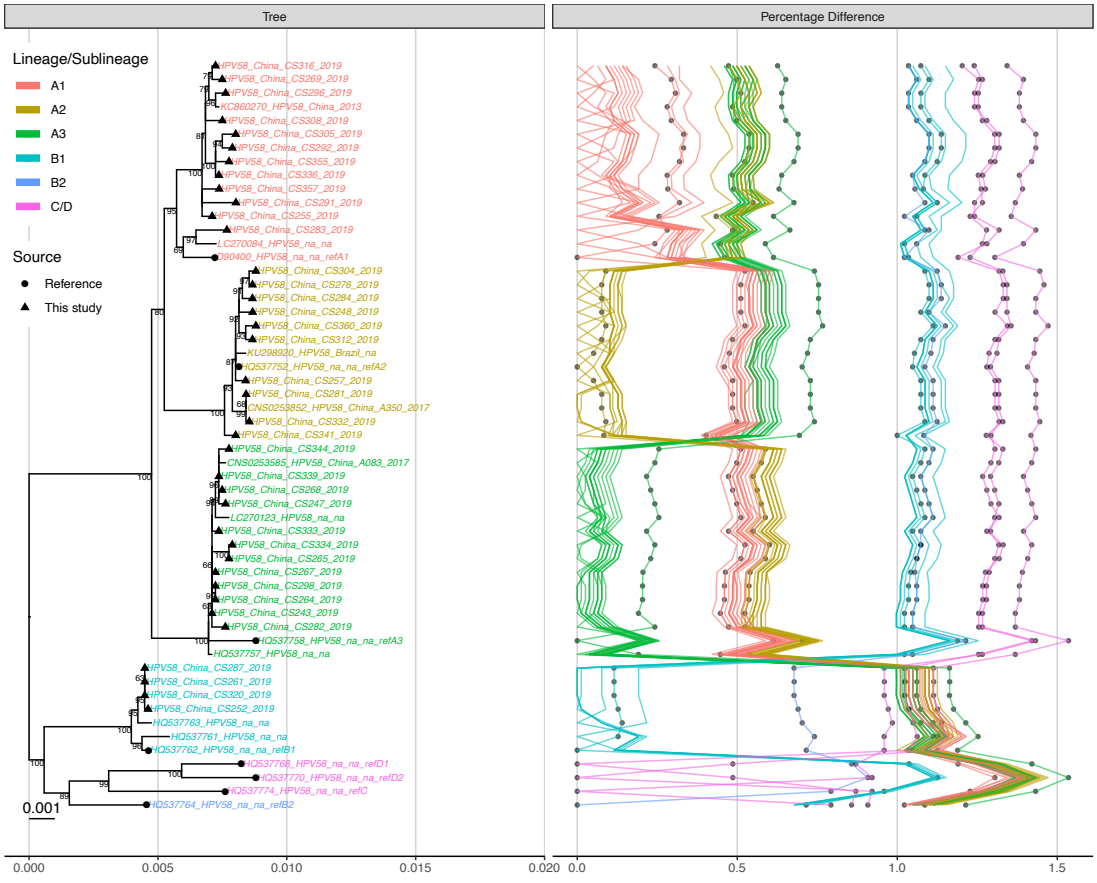
480

Figure 3：Representative Maximum Likelihood phylogeny of HPV52. Figure legends are the same as Figure 1.



485 **Figure 4：Representative Maximum Likelihood phylogeny of HPV58.** Figure

486 legends are the same as Figure 1.



487

488

Table 1：Characteristics of the HPV genomes generated by this study.

HPV type	Sublineage	Total	Clinical status				Infection type	
			HSIL	LSIL	NA	Normal	mixed	single
HPV16	A1	30	4	5	20	1	5	25
	A3	11	2	2	6	1		11
	A4	41	13	5	22	1		41
HPV18	A1	29	1	3	24	1	4	25
	A3	1			1		1	
	A4	1			1			1
	A5	1			1			1
	A6	1			1			1
	A7	2		1	1			2
HPV52	A1	5		1	4			5
	B2	104	2	7	90	5	8	96
	C2	4			4			4
	D	8			8		2	6
HPV58	A1	65	4	6	53	2	7	58
	A2	12	1	1	10		3	9
	A3	19	1		17	1	2	17
	B1	4			4			4
Total		338	28	31	267	12	32	306

NA: clinical status unknown, most probably normal or inflammation.

Table 2: Distribution of the genomic positions with conservative mutations for HPV16, 18, 52 and 58 lineages or sublineages. Blank cells indicate no detection.

HPV Type	Mutation Type	E1	E2	E5	E6	E7	L1	L2	LCR	NCR	Subtotal	Total
HPV16	missense mutation	15	25	4	8	1	13	13			79	210
	synonymous mutation	32	10	4	2	3	20	20			91	
	point mutation								33	7	40	
HPV18	missense mutation	10	21	6	1	3	13	22			76	269
	synonymous mutation	34	16	5	9	2	25	34			125	
	point mutation								31	6	37	
	deletion		6						7	18	31	
HPV52	missense mutation	12	21	7	4	8	9	5			66	289
	synonymous mutation	25	12	2	7	3	36	22			107	
	point mutation		1						37	7	45	
	deletion								19	21	40	
	insertion						3			28	31	
HPV58	missense mutation	12	3		5	6	17	7			50	186
	synonymous mutation	16	8	6	2	4	22	23			81	
	point mutation								37	6	43	
	deletion								12		12	

497 **Table 3: Unique mutations detected in certain lineages/sublineages of HPV 16,**

498 **18, 52 and 58.**

HPV_type	Sublineage	E1	E2	E5	E6	E7	L1	L2	LCR	NCR
HPV16	A4		T3524C(L257L)	A4077T(T76T)					T7177C; C7270T	
	B2						T6171G(V204G); T6685A(S375S)	A4711G(N159D)	T7328A	G4103A
	B3	A2553G(L563L)		A4062G(L71L)			C6182G(P208A)	C5239G(P335A)		
	B4	G1103C(R80T); A2601G(L579L)					G6061A(V167V)		A7232C	A4116T
	C1							A5290C(T352P)	A7839G	
	C2	T870C(A2A); C978T(N38N)	A3247G(I164M)	A4042C(I65L)	T137A(L12I)		T7132A(A524A)			
	C3			A4030C(I61L)			G5851A(R97R); A6447G(N296S)	A5544C(Q436H)		
	C4								A7688G	
	D1		G3413A(A220T)					A5073T(P279P)		
	D2		G3416A(A221T)				A6025C(T155T)			
	D3						A6803T(T415S)	T5475A(G413G)		
	D4		C3159G(T135R)						G7360A	
HPV18	A2		C3277T(A154V); T3719C(C301C)				C5769T(P114S); A6970C(K514T)	C5266T(A341A)	C7511G; A7565G	
	A3						T5619A(L64M)	T5619A(F459Y)		
	A4	A976G(Q21Q); T1452C(V180A)	G3482A(Q222Q)					C4341A(P33H); C5410T(V389V)		
	A5	C988T(D25D)	T3534C(Y240Q); C3536G(Y240Q); A3708G(S298G); T3892C(L359S)	C4085T(F50F)	A377G(E91E)		A5741G(L104L); G6089A(L220L)	C4708T(T155T); G4738A(P165P)		
	A6		G3525A(A237K); C3526A(A237K)				T5810G(S127S)			T4172G; C4210T
	A7	C1733T(L274L); G2551C(A546A)	A2972C(E52D)				C6014T(A195A); T6182G(L251L)	T4279C(A12A); C4328T(P29S); T4330C(P29S); A4366G(A41A)	T7745A	
	B1						A5924T(G165G)			
	B2	A2722C(K603N)						T4990C(I249I)		
	B3	C1362T(T150I); T1552A(A213A)	C3749A(D311E); C3902T(Y362Y)	A4151C(L72F)		C865A(N92K)	A6185G(K252K)	C4573A(S110S); T5125G(V294V)		
	C	A1636G(T241T)	G3757C(R314T); C3790T(A325V)	T3942C(S3P); T3963G(F10V)	G437A(P111P)		A5638G(N70S); G5764A(R112K); A5880G(I151V); A6224T(G265G); T6536C(N369N); A6644C(I405I); A6845C(P472P)	T4423G(G60G); G4486A(V81V); A4513T(P90P); A4726G(A161A); G5191A(R316R); C5225A(P328T); T5227C(P328T); G5351A(A370T)	T7475G	
HPV52	A2	G924A(E21K); T1469G(T202T)	A3793C(K351Q); C3837A(V365V)		A251G(L50L)		T5900C(F112F); T5999C(G145G)	T4510C(I83I)	T7744C	
	B1	C1224T(P121S); T2240C(F459F)	T3087C(Y115Y); T3403G(S221A)	T3975C(F15L); T3977G(F15L); G4154A(L74L)			A6347G(P261P); A6722C(K386N); T7121C(R519R)			
	B2	A1703G(G280G); A1877C(Q338H)	C3141T(N133N); T3281C(V180A); A3667C(T309P)				G6110A(Q182Q); T6764C(F400F); A6794G(K410K); C6824T(Y420Y)	A4480T(P73P); A5089C(A276A); T5356G(D365E)	T13C; G7168C; G7282del	
	C1	G1147A(S95N)	T3277G(C179G); C3371T(A210V); G3461A(R240Q); A3674C(K311T); T3818C(V359A)	T3973C(V14A)	C252A(R51K); G253A(R51K); G296A(M65I); T308C(F69F)		T5646C(S28P); G5864A(R100R); G6468A(G302S); A6563G(S333S)	A5147C(R296R); T5646C(V462A)	A7784C	
	C2					C662T(T37I); C733T(H61Y); T848G(L99R)			G7586A	
	D	T1222C(L120S); A1410G(S183G); A1416G(R185G); C1422T(L187L); C1622T(P253P); T1793C(S310S); C1799A(T312T); A1871T(I336I); G2042A(S393S); T2165C(D434D); A2180G(V439V); C2225T(D454D); C2351G(S496S); C2379T(L506L); A2439T(T526S); A2477T(G538G); A2537G(L558L)	T2937C(S65S); C3008A(T89K); T3022C(L94L); T3153A(I137I); A3268C(K176Q); G3289T(V183L); G3383C(C214S); G3533C(S264T); A3559G(T273A); A3579G(G279G); T3584C(V281A); C3587T(A282V); A3588G(A282V); T3639C(S299S); T3780A(R346R)	T4073G(F47L); G4088A(L52L); T4153A(L74Q)	C200T(C33C); G425T(T108T)	T573C(T7T); C766A(H72N)	C5763A(R67R); A5771T(L69L); A5873G(R103R); A5930C(P122P); T6113C(T183T); T6197C(P211P); C6483A(Q307K); T6533C(F323F); C6569A(S335S); G6659A(V365V); A6711G(S383D); G6712A(S383D); T6920G(T452T); T6935A(T457T); A6941G(Q459Q); G6980A(K472K); C6983G(D473E); T6992C(F476F); T7049C(G495G); G7079A(Q505Q)	G4366A(V35V); T4456C(S65S); T4501C(T80T); G4516A(T85T); T4609C(S116S); C4612T(G117G); T4625A(S122T); C4948T(S229S); C4975A(V238V); G5002C(Q247H); C5075G(Q272E); T5079A(L273H); T5215A(I318I); C5254A(I331I); G5305A(Q348Q); A5368G(Q369Q); G5500C(S413S)	A28G; T86C; A92C; C97T; G7168T; G7249C; G7386T; T7387G; C7571G; A7579C; TGCTGACT7700-7707del; G7713del; C7917G	G3884A; A3887C; C3889T; C3892T; C3893G; 4160_4161ins21T; GTAGATTGGCTA CATGCATAT4169-4189del; A4195C
HPV58	A2	G948A(A22A)	A2935C(S61S)	T3988C(L33L)			A6416G(R284R); T6434C(A290A)	T5143G(R300R)	C7266T	C4136A
	A3	C1965T(D361D)				C632T(T20I)	A5579C(L5F); T5747C(Y61Y)	A4935C(N231T); A5579C(M446L)	G7147T; G7194C; A7304G; A7714C; A7755G	A4192C
	B1	A2764G(I628V)	A2764G(I4M); T2953C(T6T); G3571C(V273V)		G203C(E32Q)		G5666A(E34E); G5972A(R136R); A6020G(K152K)	C4297T(Y18Y); A4498G(L85L); T4900G(S219S); G4909A(V222V); C5295A(T351N)	T7257G; A7313C; A7523del; G7619A	
	B2	T1738C(L286L)	G3562C(G270G)				T5789C(Y75Y); A6222G(I220V); G6458T(P298P); T6496C(V311A)	T5542G(T433T)	C97T; T7140G; A7435G; G7745C	
	C	T1391C(V170A); A1421C(N180T)				T852C(P93P)	C5861G(V99V); G5939A(L125L); C6038T(T158T); C6051A(P163T); A6439C(K292T); A6440C(K292T); G6450C(A296P); G6459A(D299N); T6496G(V311G); G6697A(G378D); G6711A(D383N)		TATG7185-7188del; T7189C; T7345C; T7431G	
	D1	A1068T(E62D)				G760C(G63H); G761A(G63H)		T4330C(P29P)		
	D2	A1054G(T58A)						G5266C(Q341H)	G7395A; G7421A; G7686A	G4152T

499

500

501 **Supplementary Materials**

502 **Supplementary Figure S1: Maximum Likelihood phylogeny of HPV16**. HPV
503 genomes generated by this study were combined with those from public database to
504 construct a Maximum Likelihood phylogeny with 1,000 bootstrap tests using IQ-
505 TREE. Number of genomes: 2,584. Nucleotide substitution model: GTR+F+I+G4.
506 Sources of the sequences are indicated by shapes while lineages or sublineages are
507 distinguished by different colors. Bootstrap values over 70 are labeled on nodes.
508 **Supplementary Figure S2: Maximum Likelihood phylogeny of HPV18**. Number
509 of genomes: 182. Nucleotide substitution model: TVM+F+I+G4. Figure legends are
510 the same as Supplementary Figure S1.
511 **Supplementary Figure S3: Maximum Likelihood phylogeny of HPV52**. Number
512 of genomes: 315. Nucleotide substitution model: GTR+F+I+G4. Figure legends are
513 the same as Supplementary Figure S1.
514 **Supplementary Figure S4: Maximum Likelihood phylogeny of HPV58**. Number
515 of genomes: 323. Nucleotide substitution model: TVM+F+I+G4. Figure legends are
516 the same as Supplementary Figure S1.
517 **Supplementary Table S1: Classification of the HPV16, 18, 52 and 58 genomes**
518 **obtain from Changsha.**
519 **Supplementary Table S2. Lineage/sublineage distribution of the genomes used**
520 **for mutation detection.**
521 **Supplementary Table S3. Nucleotide mutations detected in HPV16 full**
522 **genomes.**
523 **Supplementary Table S4. Nucleotide mutations detected in HPV18 full**
524 **genomes.**
525 **Supplementary Table S5. Nucleotide mutations detected in HPV52 full**
526 **genomes.**
527 **Supplementary Table S6. Nucleotide mutations detected in HPV58 full**
528 **genomes.**

References

1. Martel C de, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type: Worldwide burden of cancer attributable to HPV. Int J Cancer. **2017**; 141(4):664–670.

2. Serrano B, Brotons M, Bosch FX, Bruni L. Epidemiology and burden of HPV-related disease. Best Practice & Research Clinical Obstetrics & Gynaecology. **2018**; 47:14–26.

3. Serrano B, Sanjosé S de, Tous S, et al. Human papillomavirus genotype attribution for HPVs 6, 11, 16, 18, 31, 33, 45, 52 and 58 in female anogenital lesions. European Journal of Cancer. **2015**; 51(13):1732–1741.

4. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. **2017**; 45(D1):D499–D506.

5. Chan PKS, Ho WCS, Chan MCW, et al. Meta-Analysis on Prevalence and Attribution of Human Papillomavirus Types 52 and 58 in Cervical Neoplasia Worldwide. Zheng Z-M, editor. PLoS ONE. **2014**; 9(9):e107573.

6. Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health. **2020**; 8(2):e191–e203.

7. WHO. A Global Strategy for elimination of cervical cancer - PAHO/WHO | Pan American Health Organization [Internet]. 2018 [cited 2020 Oct 16]. Available from: http://www.paho.org/en/topics/cervical-cancer/global-strategy-elimination-cervical-cancer

8. Ginsburg O, Bray F, Coleman MP, et al. The global burden of women's cancers: a grand challenge in global health. The Lancet. **2017**; 389(10071):847–860.

9. Gultekin M, Ramirez PT, Broutet N, Hutubessy R. World Health Organization call for action to eliminate cervical cancer globally. Int J Gynecol Cancer. **2020**; 30(4):426–427.

10. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin. **2016**; 66(2):115–132.

11. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015: Cancer Statistics, 2015. CA: A Cancer Journal for Clinicians. **2015**; 65(1):5–29.

12. Xia C, Hu S, Xu X, et al. Projections up to 2100 and a budget optimisation strategy towards cervical cancer elimination in China: a modelling study. The Lancet Public Health. **2019**; 4(9):e462–e472.

13. Harden ME, Munger K. Human papillomavirus molecular biology. Mutation Research/Reviews in Mutation Research. **2017**; 772:3–12.

562 14. Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. Virology. **2013**;
563 445(1–2):232–243.

564 15. Clifford GM, Tenet V, Georges D, et al. Human papillomavirus 16 sub-lineage dispersal
565 and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-
566 positive women. Papillomavirus Research. **2019**; 7:67–74.

567 16. Mirabello L, Yeager M, Cullen M, et al. HPV16 Sublineage Associations With
568 Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200
569 Women. J Natl Cancer Inst. **2016**; 108(9):djw100.

570 17. Park JS, Shin S, Kim E, et al. Association of human papillomavirus type 16 and its
571 genetic variants with cervical lesion in Korea. APMIS. **2016**; 124(11):950–957.

572 18. Chen AA, Gheit T, Franceschi S, Tommasino M, Clifford GM. Human Papillomavirus
573 18 Genetic Variation and Cervical Cancer Risk Worldwide. Banks L, editor. J Virol.
574 **2015**; 89(20):10680–10687.

575 19. Zhang C, Park J-S, Grce M, et al. Geographical Distribution and Risk Association of
576 Human Papillomavirus Genotype 52–Variant Lineages. J Infect Dis. **2014**;
577 210(10):1600–1604.

578 20. Chan PKS, Luk ACS, Park J-S, et al. Identification of Human Papillomavirus Type 58
579 Lineages and the Distribution Worldwide. The Journal of Infectious Diseases. **2011**;
580 203(11):1565–1573.

581 21. Chen Z, Ho WCS, Boon SS, et al. Ancient Evolution and Dispersion of Human
582 Papillomavirus 58 Variants. Journal of Virology. **2017**; 91(21):17.

583 22. Yu J-H, Shi W-W, Zhou M-Y, Liu J-M, Han Q-Y, Xu H-H. Genetic variability and
584 oncogenic risk association of human papillomavirus type 58 E6 and E7 genes in
585 Taizhou area, China. Gene. **2019**; 686:171–176.

586 23. Wang R, Guo X, Wisman GBeaA, et al. Nationwide prevalence of human
587 papillomavirus infection and viral genotype distribution in 37 cities in China. BMC
588 Infect Dis. **2015**; 15(1):257.

589 24. Zhu B, Liu Y, Zuo T, et al. The prevalence, trends, and geographical distribution of
590 human papillomavirus infection in China: The pooled analysis of 1.7 million women.
591 Cancer Med. **2019**; 8(11):5373–5385.

592 25. Lu J, Shen G, Li Q, Chen X, Ma C, Zhu T. Genotype distribution characteristics of
593 multiple human papillomavirus in women from the Taihu River Basin, on the coast of
594 eastern China. BMC Infect Dis. **2017**; 17(1):226.

595 26. Sanjose S de, Quint WG, Alemany L, et al. Human papillomavirus genotype attribution
596 in invasive cervical cancer: a retrospective cross-sectional worldwide study. The Lancet
597 Oncology. **2010**; 11(11):1048–1056.

598 27. Ou Z, Chen Z, Zhao Y, et al. Genetic signatures for lineage/sublineage classification of
599 HPV16, 18, 52 and 58 variants. Virology. **2021**; 553:62–69.

600 28. Wu Y, Chen Y, Li L, Yu G, He Y, Zhang Y. Analysis of mutations in the E6/E7
601 oncogenes and L1 gene of human papillomavirus 16 cervical cancer isolates from
602 China. Journal of General Virology. **2006**; 87(5):1181–1188.

603 29. Shang Q, Wang Y, Fang Y, et al. Human papillomavirus type 16 variant analysis of E6,
604 E7, and L1 genes and long control region in cervical carcinomas in patients in northeast
605 China. J Clin Microbiol. **2011**; 49(7):2656–2663.

606 30. Tenjimbayashi Y, Onuki M, Hirose Y, et al. Whole-genome analysis of human
607 papillomavirus genotypes 52 and 58 isolated from Japanese women with cervical
608 intraepithelial neoplasia and invasive cervical cancer. Infect Agents Cancer. **2017**;
609 12(1):44.

610 31. Choi YJ, Ki EY, Zhang C, et al. Analysis of Sequence Variation and Risk Association
611 of Human Papillomavirus 52 Variants Circulating in Korea. Liu X, editor. PLoS ONE.
612 **2016**; 11(12):e0168178.

613 32. Chang Y-J, Chen H-C, Lee B-H, et al. Unique variants of human papillomavirus
614 genotypes 52 and 58 and risk of cervical neoplasia. Int J Cancer. **2011**; 129(4):965–973.

615 33. Chan PKS. Association of Human Papillomavirus Type 58 Variant With the Risk of
616 Cervical Cancer. CancerSpectrum Knowledge Environment. **2002**; 94(16):1249–1253.

617 34. Cai HB, Chen CC, Ding XH. Human papillomavirus type 16 E6 gene variations in
618 Chinese population. European Journal of Surgical Oncology (EJSO). **2010**; 36(2):160–
619 163.

620 35. Zhe X, Xin H, Pan Z, et al. Genetic variations in E6, E7 and the long control region of
621 human papillomavirus type 16 among patients with cervical lesions in Xinjiang, China.
622 Cancer Cell Int. **2019**; 19(1):65.

623 36. Yang L, Yang H, Wu K, Shi X, Ma S, Sun Q. Prevalence of HPV and variation of HPV
624 16/HPV 18 E6/E7 genes in cervical cancer in women in South West China: HPV E6/E7
625 Genes Variation in Cervical Cancer in Southwest China. J Med Virol. **2014**;
626 86(11):1926–1936.

627 37. Sun Z, Lu Z, Liu J, et al. Genetic variations of E6 and long control region of human
628 papillomavirus type 16 from patients with cervical lesion in Liaoning, China. BMC
629 Cancer. **2013**; 13(1):459.

630 38. Chang Y-J, Chen H-C, Pan M-H, et al. Intratypic variants of human papillomavirus type
631 16 and risk of cervical Neoplasia in Taiwan: HPV 16 Variants and Cervical Neoplasia. J
632 Med Virol. **2013**; 85(9):1567–1576.

633 39. Pillai MR, Hariharan R, Babu JM, et al. Molecular variants of HPV-16 associated with
634 cervical cancer in Indian population. Int J Cancer. **2009**; 125(1):91–103.

635 40. Matsumoto K. Enhanced oncogenicity of human papillomavirus type 16 (HPV16)
636 variants in Japanese population. Cancer Letters. **2000**; 156(2):159–165.

637 41. Zehbe I, Voglino G, Delius H, Wilander E, Tommasino M. Risk of cervical cancer and
638 geographical variations of human papillomavirus 16 E6 polymorphisms. The Lancet.
639 **1998**; 352(9138):1441–1442.

640 42. Boon SS, Xia C, Lim JY, et al. Human Papillomavirus 58 E7 T20I/G63S Variant
641 Isolated from an East Asian Population Possesses High Oncogenicity. Sandri-Goldin
642 RM, editor. J Virol. **2020**; 94(8):e00090-20.

643 43. Ong CK, Chan SY, Campo MS, et al. Evolution of human papillomavirus type 18: an
644 ancient phylogenetic root in Africa and intratype diversity reflect coevolution with
645 human ethnic groups. Journal of Virology. **1993**; 67(11):6424–6431.

646 44. Wang T, Zeng X, Li W, et al. Detection and Analysis of Human Papillomavirus (HPV)
647 DNA in Breast Cancer Patients by an Effective Method of HPV Capture. Banks L,
648 editor. PLoS ONE. **2014**; 9(3):e90343.

649 45. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
650 Bioinformatics. **2018**; 34(17):i884–i890.

651 46. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
652 MEM. arXiv:13033997 [q-bio] [Internet]. **2013** [cited 2020 Aug 27]; . Available from:
653 http://arxiv.org/abs/1303.3997

654 47. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: *de novo* assembly of organelle
655 genomes from whole genome data. Nucleic Acids Res. **2016**; :gkw955.

656 48. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Current Protocols in
657 Bioinformatics [Internet]. **2020** [cited 2020 Mar 6]; 69(1). Available from:
658 https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.96

659