Super-pangenome analysis of 3562 human and animal papillomavirus isolates illuminates their genome and pathogenicity evolution

Huihui Li[1^], Jingtao Chen[1^], Xinxin Zhang[1^], Xing Zhang[2], Ludong Yang[1], Yi Ding[3], Congying Chen[1], Yan Shuai[1], Ming Song[2], Jihong Liu[3], Lin Feng[1], Jing Li[1*], Jia-Xing Yue[1*]

Affiliations:

1. State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China.
2. Department of Head and Neck Surgery, Sun Yat-sen University Cancer Center, Guangzhou 510060, China.
3. Department of Gynecologic Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China.

^: contribute equally.
*: corresponding authors.

* E-mail: lijing3@sysucc.org.cn (JL); yuejiaxing@gmail.com (JXY)

Key words: Papillomavirus, HPV, pangenome, pathogenicity, genome evolution, genetic variant

Abstract

Papillomaviruses (PVs) are groups of ubiquitous DNA viruses infecting a wide range of vertebrate hosts, including human. The infection of some human PVs (i.e., HPVs) are strongly associated with cancers, but it remains elusive how such pathogenicity got evolved at the genomic level. Here we performed a comprehensive super-pangenome analysis on a compendium of well-curated 3562 human and animal PV genomes. We found their global genomic diversity being strongly associated with their body sites and tissue origins, but not with their host, geographical, or disease specificity. By reconciling time-resolved phylogenies of PVs and their hosts, we revealed a high-resolution view on how intra-host codivergence, niche adaptation, horizontal host switches, and duplication/loss turnovers collectively shaped the evolution of PVs. We identified key selection-driven mutations that defined the genus-specific divergence and pathogenic risks of PVs, with the former highly enriched in the *E1* gene and the latter showing contrasting genotype patterns between low-risk and high-risk HPVs. With independent clinical cohorts, we further demonstrated how these variants can be highly informative for HPV risk classification even at single site level, underscoring their application values in large-scale HPV screening tests. Finally, with cervical and oropharyngeal cancer patient cohorts assembled by this study, we further evaluated the prognostic value of specific HPV16 variants in predicting patient survival. Taken together, such illumination on the genome and pathogenicity evolution of PVs paves the road for better prevention, control, and treatment of PV-related cancers.

Significance Statement

Papillomaviruses (PVs) are ancient and widespread DNA viruses co-evolved with many vertebrate hosts, including human. Some human PVs (e.g., HPV16 and HPV18) bear strong association with cancers, raising a pressing public health challenge. A better understanding of how such pathogenicity got evolved holds the key in addressing this challenge. Here we systematically investigated the genomic and evolutionary diversity of PVs via a pangenome study on over 3500 human and animal PV isolates, which revealed key selection-driven mutations shaping their genome and pathogenicity evolution. With independently acquired HPV-associated cancer cohorts, we further demonstrated how some of these important HPV variants can serve as clinically informative predictive and prognostic biomarkers. Such comprehensive characterization on PVs' genome diversity, pathogenicity evolution, and clinical implication paves the road for better prevention, control, and treatment of PV-related cancers.

Introduction

Papillomaviruses (PVs) is a ubiquitous group of double-stranded DNA viruses that infect the basal keratinocytes of cutaneous and mucosal epithelia in a wide range of vertebrate host species such as fish, reptiles, birds, and mammals, including human[1]. A typical PV genome has a size of ~5–8 kb and encodes 6–9 genes[2], with genes *E5* and *E6* considered as oncogenes[3–5]. Traditionally, PVs are classified into different taxonomic groups (e.g., genera, species, and types) according to the nucleotide sequence divergence of their major capsid encoding-gene *L1*[6,7]. Members from each genus share >60% *L1* nucleotide sequence identity, while different species within a genus share 60%–70% sequence identity. A novel PV type is assigned if it shows <90% similarity to any other previously defined PV types. So far, >220 types have been defined for human PVs (i.e., HPVs), which distributed across five genera: Alphapapillomavirus (Alpha PV), Betapapillomavirus (Beta PV), Gammapapillomavirus (Gamma PV), Mupapillomavirus (Mu PV), and Nupapillomavirus (Nu PV). Interestingly, HPV types that infect mucosal epithelium was only found in Alpha PV, whereas HPVs types that infect cutaneous epithelium ubiquitously distributed across all of these five genera[1]. In comparison, PVs from animal hosts distribute across more different genera (e.g., Omicron, Upisilon, Omega, etc.) as reflected by their larger genomic divergence.

In terms of pathogenicity, many PV types are either completely asymptomatic or considered as low-risk (e.g., HPV types 6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and CP6108) as they only cause benign tumors such as warts or papilloma. In contrast, some other PVs types are classified as high-risk (e.g., HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82) or probably high-risk (e.g., HPV types 26, 53, and 66) due to their association with human cancers[6–8]. For example, it has been estimated that ~90% of cervical cancers and ~70% of oropharyngeal cancers are associated with HPV infections[9]. Therefore, a better understanding of the carcinogenicity of HPVs holds the key in the control and treatment of HPV-associated cancers.

Given the critical clinical importance of HPVs, genome sequencing and comparison have become a powerful approach to decode their genomic diversity in both human and non-human hosts. With this approach, previous studies suggested a close relationship between PVs and their host species in evolution but also noticed some violations of the pathogen-host correspondence as proposed by "Fahrenholz's rule"[10–13]. For example, PVs from five well-diverged genera (Alpha- to Nu-PVs) were all found in human, while some non-human primate PVs and HPVs are phylogenetically intermingled[14,15]. To account for such phylogenetic incongruence between PVs and their hosts, hypothesis of niche-specific adaptation within the host was proposed, suggesting PV as an ancestral generalist gradually diverged and become niche-specific specialists in their recent evolution with primate hosts (including human)[12,14,16]. In addition, sporadic cases of horizontal host transfer has also been reported, probably occurred via cross-species infection[17].

To date, there have been thousands of PV genomes available in online databases such as Genbank, with hundreds of representative ones further curated in dedicated database such as PaVE[18]. Therefore, a super-pangenome-based study that covers all currently known PV lineages isolated from both human and animal hosts is expected to provide a comprehensive view on the global genomic and ecological diversity of PVs and shed light on their host-associated evolutionary history. In this study, we systematically curated and annotated a compendium of 3562 PV genomes with diverse host species origins (e.g., human, non-human primates, non-primate mammals, reptiles, birds, and fish) to derive a high-quality super-pangenome collection for PVs. With this super-pangenome, we characterize the genomic diversity of PVs in relation to their host specificity, geographic distribution, atomical tissue origin, risk and disease association, gene repertoire, and phylogenetic clustering. More importantly, we identified key genomic variants in PV genomes underlying their genus-specific diversification and pathogenicity evolution. By further incorporating multiple independent cohorts, we demonstrated the medical relevance and application value of some key HPV variants for the precision prevention and treatment of HPV-associated cancers.

Results

A super-pangenome view on the global diversity of PVs

To capture and understand the global genomic and ecological diversity of PVs, we collected, curated, and annotated a super-pangenome collection of 3562 PV genomes via both database inquiry and literature mining (Supplementary Table 1). Thorough metadata curation was also performed regarding their host species, geographic distribution, anatomic body site and tissue origins, risk and disease associations. The 3562 PVs include 3329 HPVs, 30 non-human primate PVs, 172 non-primate mammal PVs, 6 reptile PVs, 11 bird PVs, and 14 fish PVs, covering all known host species of PVs (Figure 1A). The 3329 HPVs collected here were isolated from the global population (Figure 1B), with most of them classified into the Alpha (84.9%), Beta (2.8%) and Gamma (12.2%) genera (Supplementary Table 1).

These 3562 PV genomes show rich variation in their genome content, genome size (4.7 kb–10.4 kb) and GC% (35.1%–59.2%). Regarding genome content, the *E1*, *E2*, *L1*, and *L2* genes are universally conserved and constitute the core gene set of PVs, whereas the other genes exhibited variable presence/absence patterns (Supplementary Figure 1). As for genome sizes, fish PVs have clearly smaller genome sizes, largely due to their general depletion of the *E5*, *E6*, and *E7* genes (Figure 1C). Within the 3329 HPVs, significantly larger genome sizes were noticed for Alpha-PVs, driven by expanded noncoding regions, especially the URR region that regulates viral transcriptional activity (Supplementary Figure 2). The extreme genome sizes of two HPV outliers are caused by structural variants (segmental deletion for HPV16_K2821 and segmental duplication for HPV11_GUMC_AJ_Lung), with the latter one experimentally verified in the original study [19] (Supplementary Figure 3). Regarding GC%, we observed a significant positive correlation between the genomes of PVs and their hosts both before and after adjusting for their respective phylogenetic relatedness (see Methods; Supplementary Figure 4). We systematically compared individual codon usage frequencies of PVs and their hosts in 41 PV-host pairs covered by our dataset and found 87.8% (36/41) of them showing significant positive correlations (Supplementary Figure 5). This observation hints for a potential selection on host-matching codon usages for PVs to better exploit the host translation machinery for completing their viral life cycles. Interestingly,

such positive correlation seems diminished for HPVs, especially for high-risk/probably-high-risk ones. Our examination on codon adaptation index (CAI), a measurement for host-dependent codon usage adaptation, provides further supports for this pattern, suggesting potentially strong counter selection force from the human immune system against HPVs with high pathogenicity (Figure 1D).

We constructed all-against-all whole genome alignment for 3548 PV genomes from birds and terrestrial vertebrates and used the HPV type 16 (HPV16) as the reference to chart global genomic variation landscape of PVs with extracted single nucleotide variants (SNVs). The fish PVs were excluded for this analysis due to their large genomic divergence from other PV genomes. Principal component analysis (PCA) shows clear genomic differentiation between Alpha-PVs and PVs from other genera (Figure 1E). Such bipartition also holds in terms of anatomic body sites and tissue origins but not so much regarding host specificities, disease status, or geography (Figure 1F, G; Supplementary Figure 6). A clear distinction between high-risk and low-risk HPVs was further captured in our PCA analysis (Figure 1H). These observations collectively suggest substantial genomic divergence have taken place during the rise of Alpha-PVs as well as those high-risk HPVs, which likely opened opportunities for their niche expansion and pathogenesis.

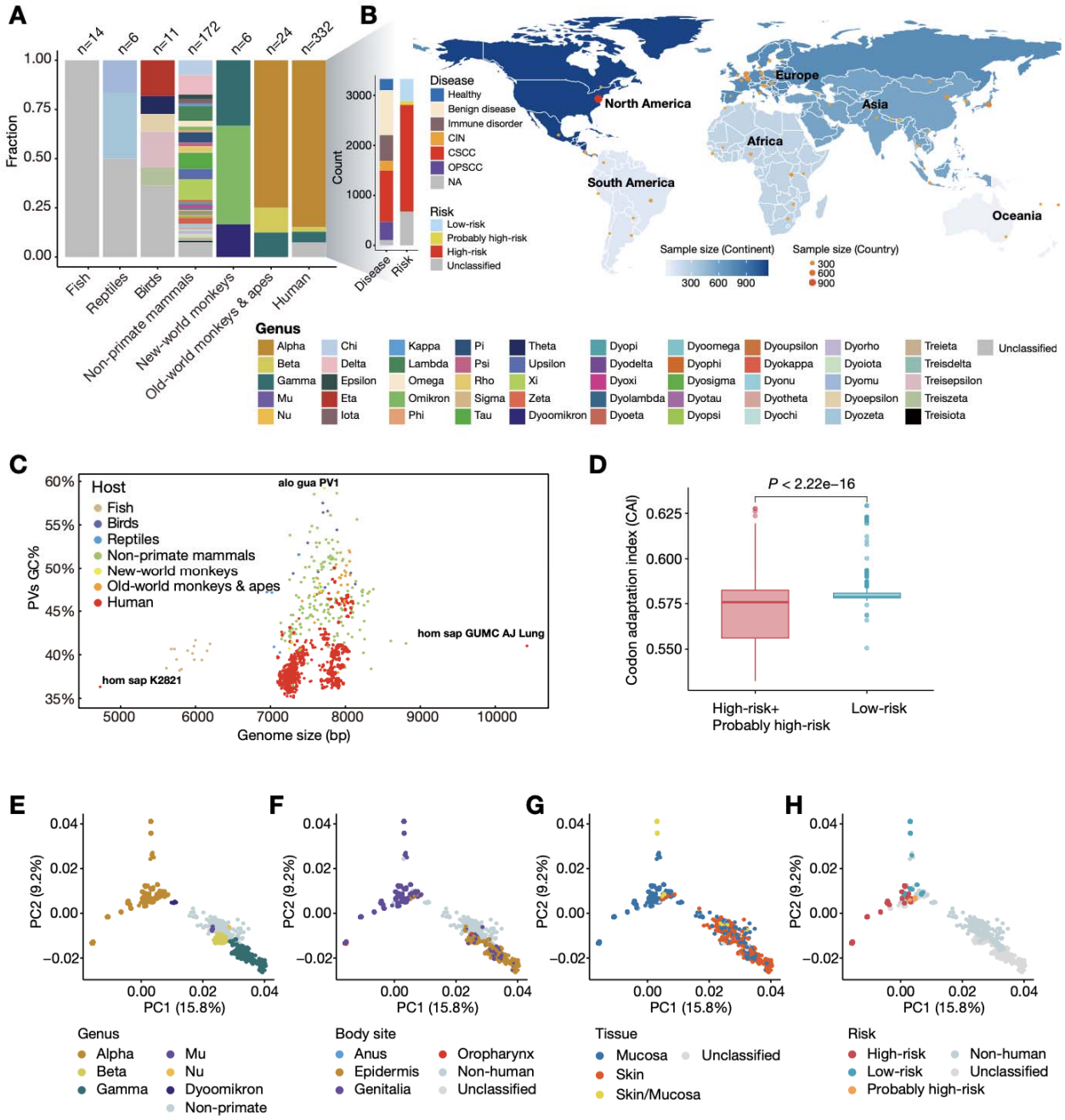


Figure 1. The global sampling and genomic diversity of the 3562 PV genomes. (A) The host and genus distribution of the 3562 PV genomes. (B) The geographical, disease type, and risk classification of the 3329 HPV genomes. (C) The genome size and GC% distribution of the 3562 PVs with regard to the host taxonomy groups. (D) The positive PV-host correlation in GC%. (E-H) The principal component analysis of the global PV genome diversity with regard to the PV genus classification (E), anatomic body site origin (F), anatomic tissue site origin (G), and high-risk vs. low-risk pathogenicity classification (H).

Time-resolved phylogeny of PV evolution

Our phylogenetic analysis on 3548 bird and terrestrial vertebrate PV genomes revealed a well-resolved phylogeny with 36 pure and 33 mixed clades (Supplementary Figure 7). Downsampled based on this phylogeny, we further generated a representative subset containing 600 PV genomes (including all 219 animal PVs and 381 representative HPVs covering all HPV lineages) to reduce redundancy in some heavily sequenced HPV types and built the corresponding downsampled PV tree (Figure 2 and Supplementary Figure 8). The 3548-genome-based and 600-genome-based PV trees show highly consistent topology, with PVs designated as the same type (e.g., HPV16) always form the same clade, while PVs designated with the same genus (e.g., Alpha, Beta, etc.) predominantly clustered together as well. With the 600-PV-based tree, we evaluated the phylogenetic congruency between genome tree with gene trees derived from individual gene (e.g., *E1*, *E2*, *L1*, *L2*, etc.) or paired gene combinations (e.g., *E1-E2*, *E1-L1*, etc.) (Supplementary Figure 9, 10; Supplementary Table 3). The phylogenetic distances between the genome tree and gene tree negatively correlates with the sizes of the corresponding genes or gene pairs, with the gene pair E1-L1 showing highest consistency with the genome tree. When only considering single genes, E1-based gene tree better recaptured the genome-level phylogenetic signals compared with the more widely used L1-based gene tree.

Next we performed Bayesian molecular dating analysis on the 600-PV-based downsampled dataset to establish the evolutionary time frame of PVs (Figure 2 and Supplementary Table 4). We estimated the divergence between the Alpha- and Beta/Gamma-super clades occurred at 96.7 million years ago (MYA). Given that Alpha-PVs are mostly mucosa-related whereas Beta/Gamma-PVs are predominantly skin-related, one can conjecture that such novel niche adaptation to mucosa is likely an important driver underlying such early divergence. The separation between the Beta- and Gamma-super-clades occurred more recently at 80.7 MYA. Interestingly, two bat PVs (min_sch_PV1 from the little bent-wing bat and ept_ser_PV2 from the serotine bat) branched off immediately before the Beta-Gamma divergence, hinting for the potential role of bats in mediating the early divergence of PVs. We noticed that the non-primate to primate host switch occurred within these three super clades roughly around the same time at 69.5-75.1

MYA, indicating a scenario of concordant non-primate-to-primate infection after the intra-host early divergence in non-primate mammals. Considering that the divergence time between human and the most closely related non-primate mammal flying lemur is 70.5-86.2 MYA, the primate PV infection likely occurred vertically along with the rise of primates. Within the primate PVs, the Alpha-PVs originated around 44.9 MYA, in sync with the separation of apes and old-world monkeys from new-world monkeys.

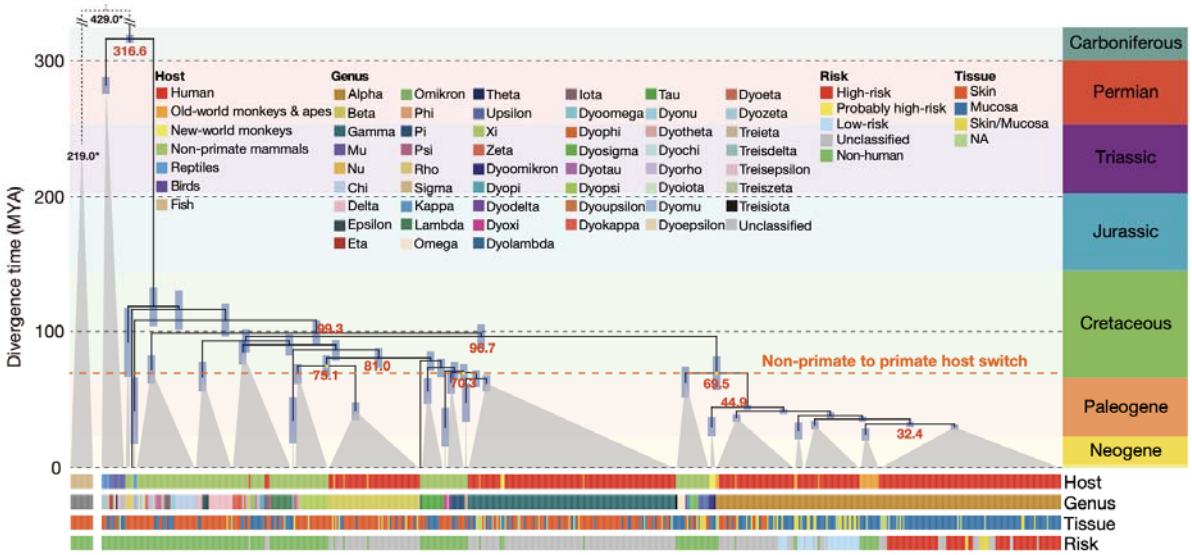


Figure 2. Evolutionary time frame of PV genome divergence and evolution. Calibrated time-resolved evolutionary tree for 600 representative PV genomes, with the branch lengths proportional to divergence time in million years ago (MYA). Key divergence time points regarding the early divergence of PVs were denoted in red, with the red dash line further highlighted concordant host shifts from non-primate mammals to primates for PVs from the Alpha, Beta, and Gamma super clades. The shades in each node represent 95% highest posterior density (HPD) interval. The time tree of fish PVs was further annexed by referencing to their host divergence time (denoted with asterisk). The tracks at the bottom show the phylogenetic distribution of PVs regarding their host species group, genus classification, tissue origin, and risk classification.

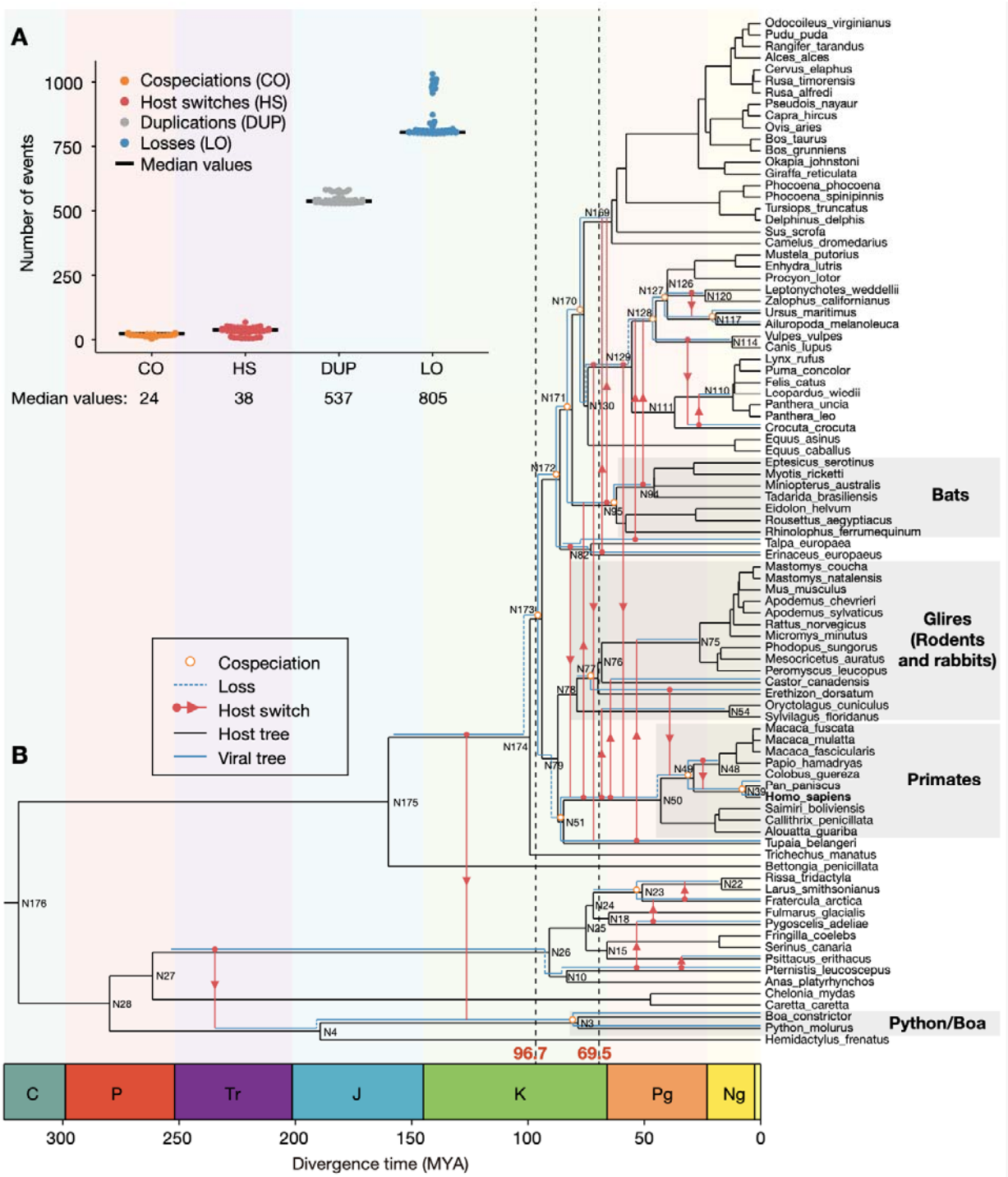


Figure 3. Phylogenetically reconciled evolutionary events during PV evolution. (A) The number of different types of evolutionary events inferred under a range of event-specific cost parameter settings. CO: cospeciation, HS: host switch, DUP: duplication; LO: loss. (B) Horizontal host switch events revealed by reconciliation between the virus and host phylogenetic trees. The phylogenetic node IDs associated with these host switch events were indicated, which were referred to in Supplementary Table 6. Geologically defined time zones are further depicted underneath. C: Carboniferous; P: Permian; Tr: Triassic; J: Jurassic; K: Cretaceous; Pg: Paleogene; Ng: Neogene.

Key divergence time point such as 96.7 MYA for Alpha vs. Beta/Gamma PVs and 69.5 MYA for non-primate to primate host switch were also indicated.

While PVs generally followed a virus-host co-divergence model, occasional instances of virus-host phylogenetic incongruency have also been reported [13,20]. Our analysis recaptured these observations, hinting for sporadic horizontal host switches. For example, some HPVs such as HPV1, HPV63, and HPV41 directly clustered with non-primate mammalian PVs from beaver and porcupine, while PVs from boa and python are deeply embedded in mammalian PVs rather than clustered with PVs from reptiles and birds. We employed a co-phylogenetic statistical framework to systematically assess such virus-host phylogenetic incongruence and to infer optimal evolutionary scenarios for reconciliation (Figure 3 and Supplementary Table 5,6). Our results revealed a consistently high level of duplication/loss turnover in PV genome evolution regardless of event-specific cost parameter settings (Figure 3A). We detected frequent horizontal host switch events involved with bats (13.6%), rodents/rabbits (18.2%), and primates (31.8%) (Figure 3B). For example, we found bats are strong candidates as the original source of HPV41 and its relative in porcupine, which are further nested within a bat-specific PV subclade that contains eight PVs from five bat species, hinting that bats should be the original source. Such a refined view on both when and how these host switch events occurred provides a new foundation for future investigation with enhanced PV sampling in these taxa.

Selection driven mutations underlying the diversification among Alpha-, Beta-, and Gamma-PVs

Given the observed early divergence among Alpha-, Beta-, and Gamma-PVs, we set out to identify key mutations that define their genomic diversification. Following the phylogenetic tree, we screened for SNVs sites that got progressively enriched in the Beta- and Gamma-super-clades as well as those got progressively depleted in the Alpha-super-clade (because the HPV16 reference genome used in our SNV calling belongs to the Alpha-super-clade). We postulated that these SNV sites, whether in their derived or reference states, should represent the genomic diversification of the Alpha-, Beta-, and Gamma-PVs. In this way, we identified 29, 54, and 15 lineage-defining SNVs for the Alpha-, Beta-, and Gamma-super-clades respectively, along with another 11 SNVs shared by both Beta- and Gamma-super-clades (Figure

4A and Supplementary Table 7). Evaluated by mixed effects model of evolution (MEME) [21], we found 15 of these super-clade-specific SNVs (corresponding to 12 amino acid substations in L1 and E1) under episodic diversifying selection (MEME test, $P<0.1$) (Figure 4B and Supplementary Figure 11). The fact that those with strong diversifying selection signals (i.e., MEME test $P<0.05$) are either Alpha-specific (n=5) or Beta-and-Gamma-specific (n=2) highlights the role of such site-specific selection in driving the early divergence of PVs (i.e., Alpha vs. Beta/Gamma). By using the bird and reptile PVs (clade M1) as the outgroup, we further determined the directions of these selection-driven mutations in evolution (Figure 4B and Supplementary Figure 11). Notably, the *E1* gene was strongly selected during the rise of Alpha PVs, as all but one selection driven mutation occurred in *E1*, especially its encoded N-terminal regulatory domain (Figure 4C). Given that a major difference between Alpa PVs and Beta/Gamma PVs lies in their infected tissue type (mucosa vs. cutaneous epithelium), it is very possible that these N-terminal mutations contribute to such tissue-specific adaptation. The remaining selection-driven mutation that defines the Alpha PV lineage locates in the *L1* gene, with its particular mutational site (L1: p.430 T->K) covered by an experimentally validated epitope (L1: p.427–445) [22,23] (Figure 4D). Future studies on these selection-driven lineage-specific mutations are expected to fully elucidate their functional and evolutionary impacts, especially in the context of Alpha PV evolution.

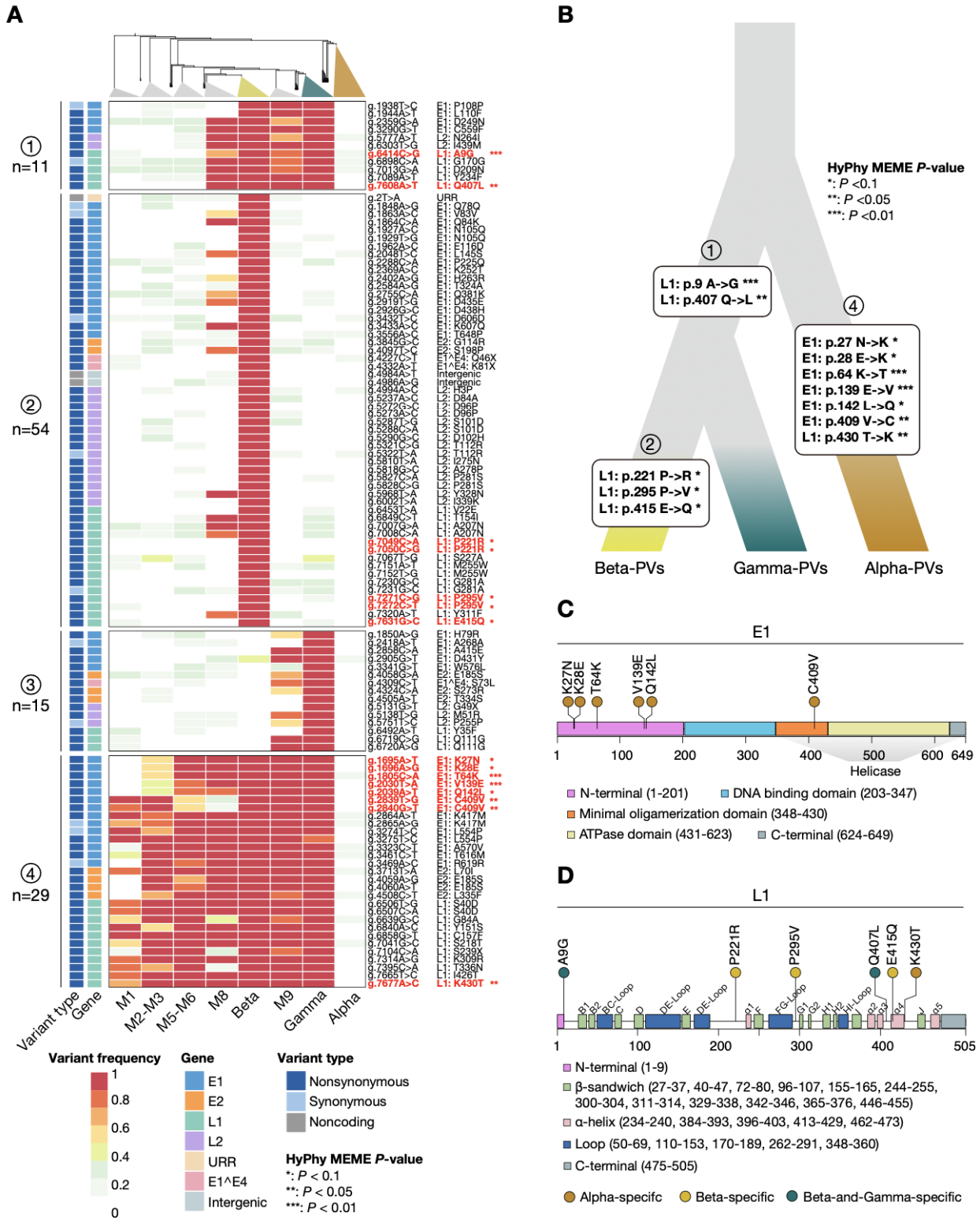


Figure 4. Key mutations defining the early divergence of the Alpha-, Beta-, and Gamma-PVs. (A) The prevalence heatmap of identified key mutations got progressively enriched in the Beta-and-Gamma-super-clades (②),

Beta- (▯), and Gamma-super-clades (▯), or progressively depleted in the Alpha-super-clades (▯). The genome and protein coordinates of these mutations were listed on the right, with the red fonts indicating mutations under significant diversifying selection (**P*<0.1, ***P*<0.05, ****P*<0.01). (B) The phylogenetic distribution and mutational directions of key mutations under diversifying selection that drove the super-clade-specific diversification.

Risk-defining HPV variant sites separating low-risk and high-risk HPVs

Our pangenome analysis reveals a highly correlated phylogenetic structure of HPVs regarding their pathogenicity classifications (Figure 5A, B). All "high-risk" and "probably high-risk" HPVs tightly clustered into a single derived lineage within the Alpha-super-clade, suggesting a lineage-specific origin of high pathogenicity. Referring to our molecular dating analysis, we estimated these high-risk HPVs arose around 32.4 MYA, soon after the divergence between human and old-world monkeys. The two most common high-risk HPVs, HPV16 and HPV18, got diversified at 1.4 MYA and 0.6 MYA respectively, suggesting a scenario of ancient high-risk HPV diversification (both across and within different HPV types) in archaic hominins, before the rise of modern human [24].

To identify key mutations associated with the rise of high-risk PVs, we first evaluated the branch-specific selection scheme around this high-risk-defining branching point based on adaptive branch-site random effects likelihood (aBSREL) for different HPV genes [25]. Evidence of episodic diversifying selection (aBSREL test, $P<0.05$) in *E1* and *E2* genes were detected immediately after the branching point (Supplementary Figure 12). Since *E1* and *E2* are essential for HPVs' viral genome replication, it is no surprising that their coding sequence was positively selected along with the rise of high-risk HPVs. At site-specific level, we found a total of 34 nonsynonymous SNV sites as well as 6 synonymous SNV sites showing contrasting genotypes between low-risk and high-risk HPVs (Figure 5C, Supplementary Figure 13, and Supplementary Table 8). We define these 40 SNV sites as risk-defining signature sites that are highly informative in differentiating high-risk HPVs from their low-risk counterparts.

Half (21/40) of these risk-defining signature sites locate in *E1* and *E2* genes, with their genomic distribution peaks right next to known HPV host-integration breakpoint hotspots[26] (Figure 5D). Also,

we found these signature SNV sites tend to be in close adjacency to, or even directly overlapped with, previously characterized HPV-human microhomology sequences, which themselves promote the host genome integration for high-risk HPVs[27] (Supplementary Figure 14). These observations collectively suggest that the contrasting mutational patterns at our identified signature sites between low-risk and high-risk HPVs might be intrinsically linked to the microhomology-mediated DNA repair process accompanied with high-risk HPVs' host genome integration. In addition to mutation, natural selection likely also played an important role in maintaining diversified genotypes at these sites between low-risk and high-risk HPVs. We detected site-specific diversifying selection (MEME test, $P<0.1$) at five nonsynonymous sites (four in *E1* and one in *E2*) of our identified signature SNV sites. Among them, the four *E1* SNV sites all locate in E1's helicase domain while the remaining *E2* SNV site falls in E2's transactivation domain, both of which are critical domains impacting HPVs' viral replication activity. Changes in the intramolecular interactions of E1 and E2 were also respectively noted for four out of five of these sites at the structural level (Supplementary Figure 15). Aside from such direct functional impacts on E1 and E2, specific genotypes selected at these sites could also contribute HPVs' pathogenicity by promoting immune evasion. Experimentally validated immune epitopes can be found for three (two in *E1* and one in *E2*) of these high-risk-defining variants and we found the epitopes carrying the high-risk variant forms generally exhibit lower affinity to major histocompatibility complex class II (MHC-II) molecules compared with those carrying the low-risk variant forms (Figure 5E-G and Supplementary Table 9, 10). This finding is concordant with previous reports from genome-wide association studies (GWAS) showing that MHC-II genes are strongly associated with the subjectivity of cervical and oropharyngeal cancers, two most notable HPV-associated cancer types [28–30]. Therefore, the selected genotypes at these signature SNV sites in high-risk HPVs likely confer them with stronger escaping potentials from the host immune surveillance.

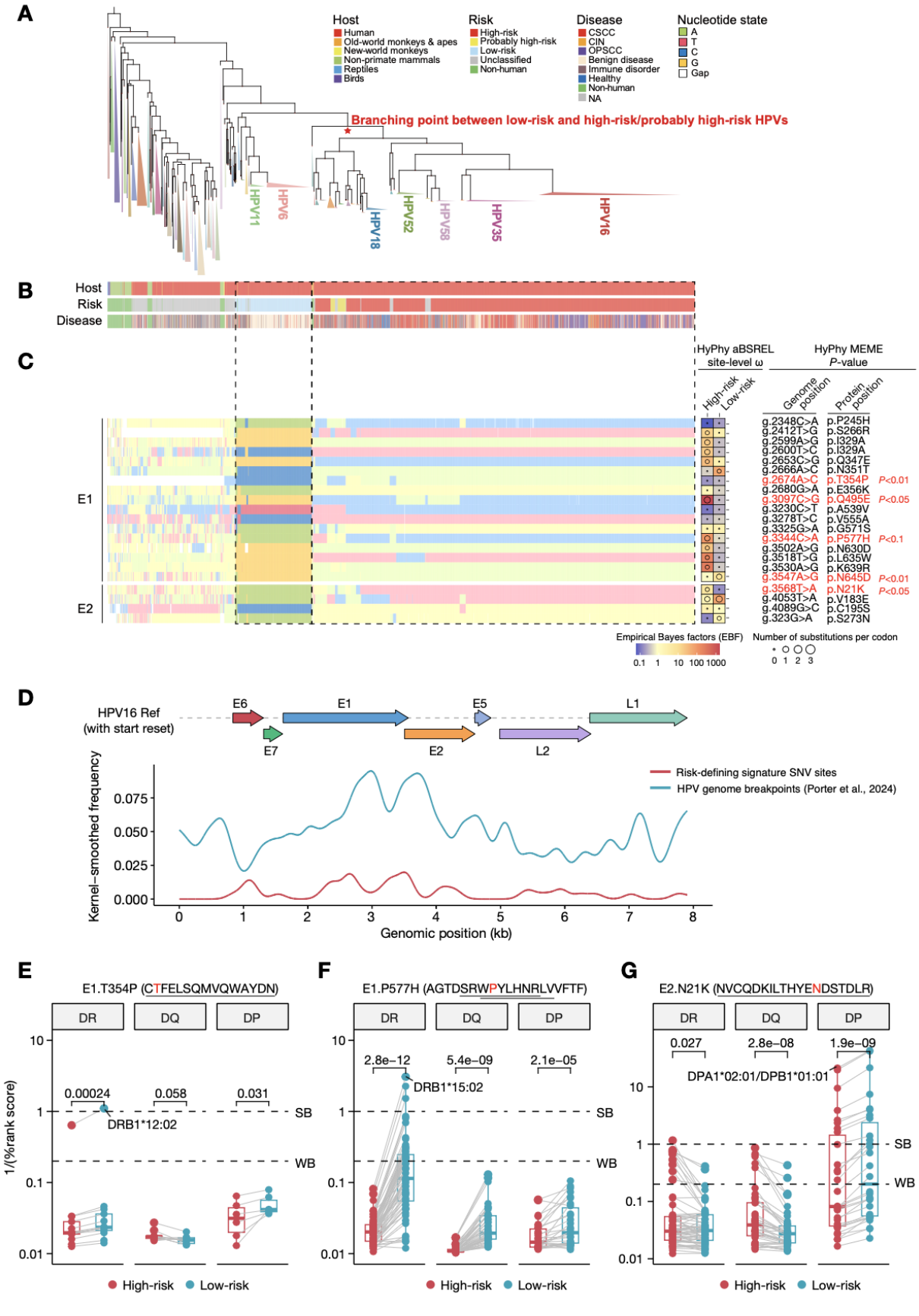

A
Host
Human
Old-world monkeys & apes
New-world monkeys
Non-primate mammals
Reptiles
Birds
Risk
High-risk
Probably high-risk
Low-risk
Unclassified
Non-human
Disease
CSCC
CIN
OPSCC
Benign disease
Immune disorder
Healthy
Non-human
NA
Nucleotide state
A
T
C
G
Gap
Branching point between low-risk and high-risk/probably high-risk HPVs
HPV11
HPV6
HPV18
HPV52
HPV58
HPV35
HPV16
B
Host
Risk
Disease
C
HyPhy aBSREL site-level ω
HyPhy MEME P-value
High-risk
Low-risk
Genome position
Protein position
E1
E2
g.2348C>A p.P245H
g.2412T>G p.S266R
g.2599A>G p.I329A
g.2600T>C p.I329A
g.2653C>G p.Q347E
g.2666A>C p.N351T
g.2674A>C p.T354P P<0.01
g.2680G>A p.E356K
g.3097C>G p.Q495E P<0.05
g.3230C>T p.A539V
g.3278T>C p.V555A
g.3325G>A p.G571S
g.3344C>A p.P577H P<0.1
g.3502A>G p.N630D
g.3518T>G p.L635W
g.3530A>G p.K639R
g.3547A>G p.N645D P<0.01
g.3568T>A p.N21K P<0.05
g.4053T>A p.V183E
g.4089G>C p.C195S
g.323G>A p.S273N
Empirical Bayes factors (EBF)
0.1 1 10 100 1000
Number of substitutions per codon
0 1 2 3
D
HPV16 Ref (with start reset)
E6
E7
E1
E2
E5
L2
L1
Kernel−smoothed frequency
0.075
0.050
0.025
0.000
Risk-defining signature SNV sites
HPV genome breakpoints (Porter et al., 2024)
0 1 2 3 4 5 6 7 8
Genomic position (kb)
E
E1.T354P (CTFELSQMVQWAYDN)
DR
DQ
DP
1/(%rank score)
10
1
0.1
0.01
0.00024
0.058
0.031
DRB1*12:02
SB
WB
High-risk
Low-risk
F
E1.P577H (AGTDSRWPYLHNRLVVFTF)
2.8e−12
5.4e−09
2.1e−05
DRB1*15:02
G
E2.N21K (NVCQDKILTHYENDSTDLR)
0.027
2.8e−08
1.9e−09
DPA1*02:01/DPB1*01:01

Figure 5. Risk-defining signature SNV sites with contrasting genotypes between low-risk and high-risk HPVs. (A) The phylogenetic tree of 3548 PVs. (B) The metadata track for the 3548 PVs regarding their species origin, pathogenic risk, and associated diseases. (C) The nucleotide state of risk-defining signature SNV sites in *E1* and *E2*. Sites highlighted in red are subject to significant diversifying selection evaluated by HyPhy MEME. (D) The physical distribution of risk-defining SNV sites and the breaking points of the HPV genome. (E-G) The MHC-II binding affinity comparison of epitopes between high-risk and low-risk epitopes in E1 (E, F) and E2 (G). The amino acid changes in epitope peptides (underlined sequences) were indicated above, while the MHC-II binding affinity [measured by 1/(%rank_score)] for the top 3 MHC-II alleles from nine geographical regions were plotted below. The MHC-II allele with the highest binding affinity were highlighted. The *P*-value of Wilcoxon signed-rank test (paired, two-sided) is given. Dashed lines represent the thresholds for strong binding (SB, 1%) and weak binding (WB, 5%).

Clinical implications of HPV variants in pathogenic risk and prognosis prediction

Conventionally, the pathogenetic risk of an HPV isolate is scored based on its specific subtype (e.g., high risk for HPV16 and low risk for HPV6). Given the strong segregating patterns between low-risk and high-risk HPVs in our identified risk-associated signature SNV sites, it could be possible to predict high-risk vs. low-risk HPVs by just knowing the genotypes of these sites. Here we systematically evaluated their prediction value in pathogenic risk prediction with four independent cohorts, which include 1) 194 HPVs genomes annotated as “complete” in Genbank, 2) 5871 HPV genomes annotated as “partial” in Genbank, 3) 131 HPV genomes sequenced from HPV16-positive cervical cancer (cervical squamous cell carcinoma, CSCC) patients examined in this study, and 4) 63 HPV genomes sequenced from HPV16-positive oropharyngeal cancer (oropharyngeal squamous cell carcinoma, OPSCC) patients examined in this study (Supplementary Table 11-14). We verified the HPV genomes that we assembled with cohort 3 and 4 are indeed HPV16 genomes based on a phylogenetic analysis covering all of them as well as the 600 PV set (Supplementary Figure 16). For each of our identified 34 nonsynonymous risk-defining SNV sites, we examined how well could our identified representing genotypes for high-risk or probably high-risk HPVs in our pangenome analysis be recovered by the high-risk or probably high-risk HPVs included in the corresponding validation cohort (Figure 6 A, B and Supplementary Table 15-18). Here we used a set of widely used classification metrics such as precision, recall, and F_1 score for this benchmark. For cohort 1 and 2, both low-risk and high-risk HPVs are included, therefore all these three metrics can be readily

calculated. Regardless which metric was used, we found 27 (for cohort 1) and 24 (for cohort 2) of the 34 signature sites exhibit satisfying performance scores >=0.9, indicating well-balanced sensitivity and specificity in HPV risk classification even just based on just a single SNV site. For cohort 3 and 4, since only high-risk HPV genomes (HPV16) were included, we calculated the recall value for each signature SNV site. For almost all sites (32/34 for CSCC and 33/34 for OPSCC), we got perfect recall values of 1.0 with these two cohorts, which again highlights the application value of our identified risk-associated signature SNV sites in predicting HPVs' pathogenic risk.

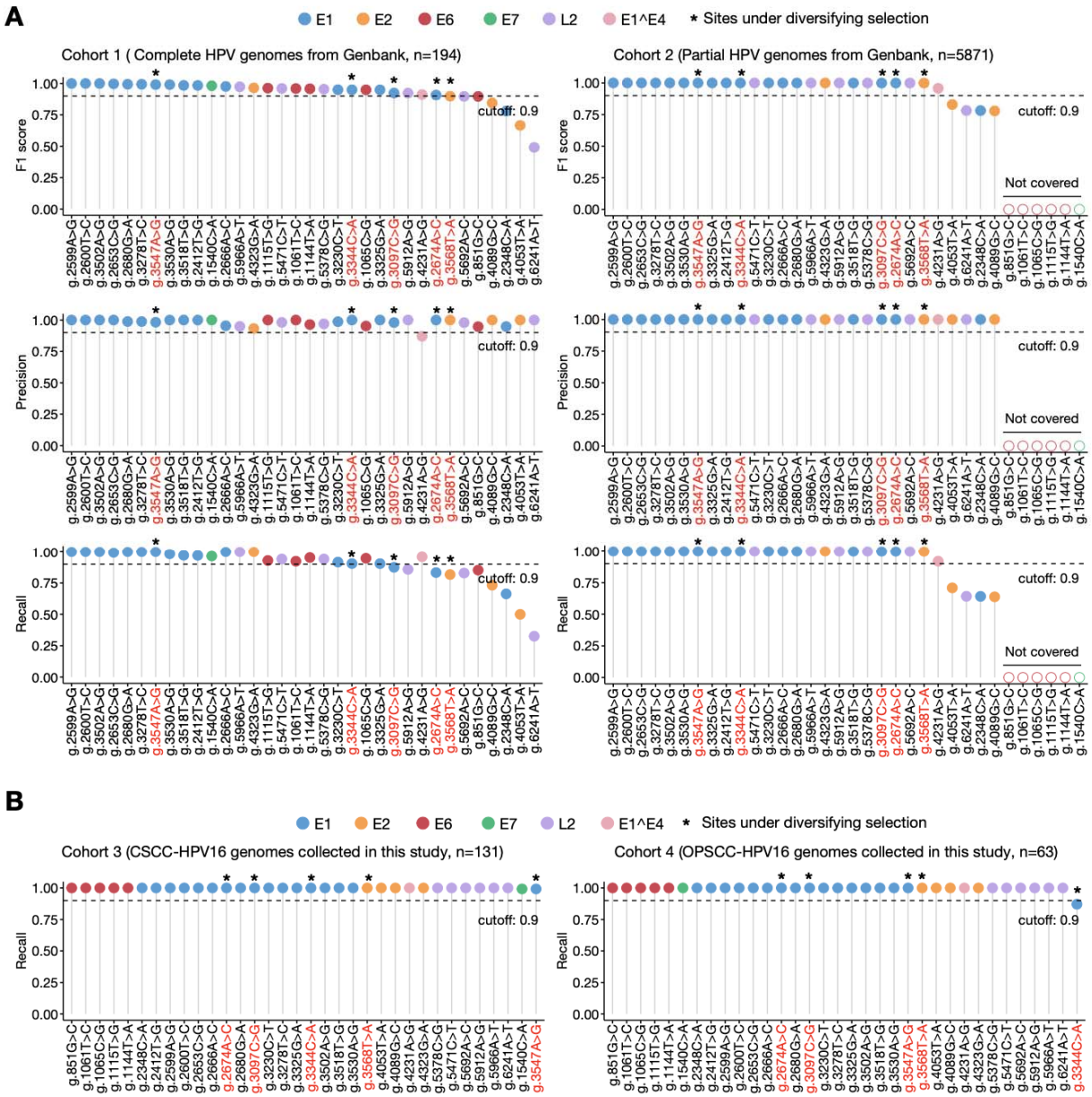


Figure 6. Performance validation of signature HPV SNV sites for pathogenic risk prediction. (A) The prediction performance of each individual risk-associated signature site in identifying HPVs with high-risk or probably high-risk pathogenicity in cohorts retrieved from Genbank (left: cohort 1; right: cohort 2). (B) The prediction performance of each individual risk-associated signature site in identifying HPVs with high-risk or probably high-risk pathogenicity in cohorts collected in this study (left: cohort 3; right: cohort 4). The five selection-drive sites are further highlighted in red color as well as with the asterisk notation. The genomic coordinates indicated are based on the start-reset-version of the HPV16 reference genome.

In addition to risk prediction for HPVs, the identification of biomarkers that predict patient survival is also of importance in clinical practice. For HPV-associated cancers, the positivity of high-risk HPVs (e.g., HPV16) has been used as a primary prognostic predictor in this regard[31–33]. Until recently, much less is known regarding the prognostic value of specific variants in high-risk HPV genomes on their associated cancer patients[34]. With the HPV16-positive CSCC (n=131) and OPSCC (n=63) patient samples collected in this study (cohort 3 and cohort 4 aforementioned; mean follow-up time: 5.39 year for CSCC and 5.08 year for OPSCC), we identified four HPV16 variants (three for CSCC and one for OPSCC) with significant impacts on patient overall survival (OS) after the false discovery rate (FDR) correction (Figure 7A-D, Supplementary Figure 17, and Supplementary Table 19-20). The three CSCC-prognostic variants (two in L2 and one in L1) show strong linkage to each other, with univariate hazard ratios ranging from 10.34 to 15.46 (Supplementary Figure 17). As for OPSCC, our identified one prognostic variant locates in E2 with a univariate hazard ratio of 22.71. Multivariate Cox regression analysis on these four prognostic HPV16 variants suggested added predictive accuracy in addition to traditional prognostic factors such as age and tumor stage (Figure 7E). Two (L2 p.306 and E2 p.198) of these four prognostic variants are nonsynonymous, and they both show notable changes in intramolecular interactions at the structural level (Figure 7F, G and Supplementary Figure 18). Taken together, although with relatively small cohort size, our analysis provides a proof of concept that some HPV variants could have significant impact on the survival of HPV-associated cancer patients. Future studies with larger human cohorts from distinct ethnic groups are expected to reveal more such variants that could be used collectively towards a better prognosis prediction.

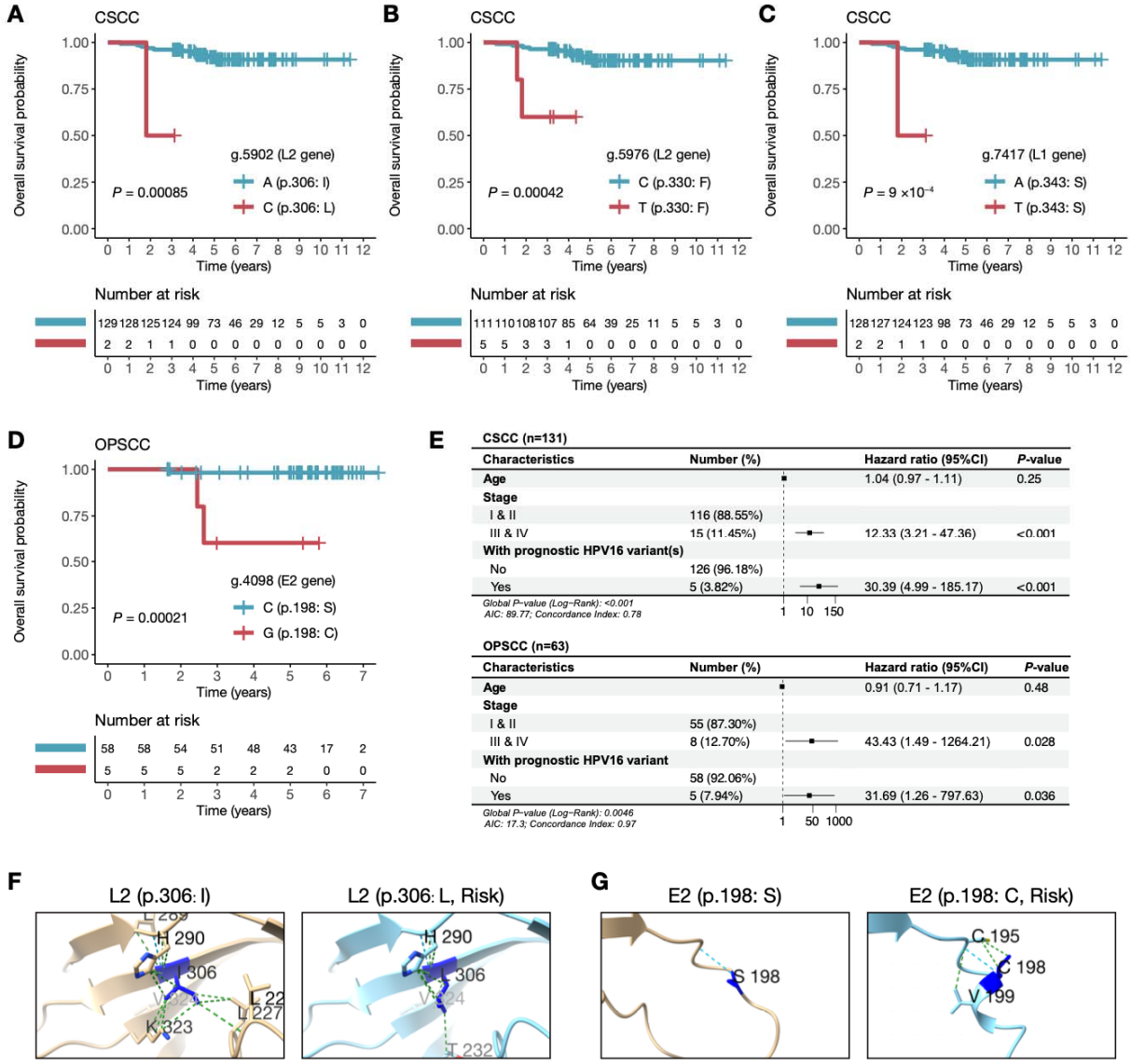


Figure 7. Discovery of prognosis-associated HPV16 variants. (A-D) The overall survival (OS) of patients stratified by individual prognostic HPV16 variants identified in our CSCC (A–C) and OPSCC (D) cohorts. (E) Prognostic factors for overall survival (OS) by multivariate analysis. (F-G) The structural view of the two nonsynonymous prognostic variants identified in *L2* (F) and *E2* (G) and their impacts on intramolecular interactions.

Discussion

PV is an ancient virus that have evolved over millions of years and infected a wide range of animal hosts, including human. With such long-term virus-host co-evolution, it is theoretically expected that the pathogenicity of the virus tends to mitigate to reach a state of chronic inapparent infection, especially for the slowly evolved double stranded DNA viruses such as PVs[35]. Indeed, this seems to be the case for PVs evolved in most animal lineages. However, specific types of human PVs (i.e., HPVs) such as HPV16 and HPV18 have been proven to be highly carcinogenic, imposing a global public health threat to human society. Therefore, it is important to study the genetic and molecular basis of such pathogenicity evolution. The rapid progress of sequencing technologies and their increasing adoption in HPV-positive patients and animal hosts in recent years vastly expanded the genomic resources of PVs available[18,36], enabling a comprehensive comparative study between low-risk and high-risk HPVs via the lens of genome evolution.

In this study, we systematically collected and curated 3562 human and animal PV genome sequences to generate a super-pangenome collection of PVs covering its global diversity. Using comparative genomics and molecular evolution frameworks, we established a pangenome view for PVs genome evolution, revealing their genome size and codon usage preference changes along the rise of Alpha PVs with high pathogenicity. We identified clear genomic differentiation between Alpha PVs and PVs from other genera is correlated with their unique body parts and tissue specificity, underscoring its implication in PVs' niche adaptation. Especially, we found sites under such lineage-specific selection predominantly fall in the N-terminal regulatory domain of the E1 protein, highlighting the importance of this domain in promoting Alpha PVs' niche expansion regarding both body parts and tissue types. Indeed, previous studies found this N-terminal regulatory domain of E1 can interact with the cellular WD repeat protein p80 (WDR48) from anogenital HPV but not with cutaneous HPV or animal PVs, therefore linking HPVs' N-terminal domain with their niche specificity[37,38].

As most viruses rely on the translation machinery of their hosts, tRNAs genes in particular, for efficiently completing their viral life cycles, a similar codon usage between a virus and its host is expected to be

beneficial[39]. With some notable exceptions (as also observed in this study), the phylogenetic distribution of PVs and their hosts generally matched with each other, suggesting a relatively strong pattern of host-specific association for millions of years. Therefore, one would expect PVs to co-evolve with their hosts regarding codon usage preference. Indeed, this is what we observed for a great majority of the cases (87.8%). However, to our surprise, HPVs' codon usage preferences seem to be considerably different from that of human, especially for HPVs classified as high-risk or probably high-risk. It is tempting to hypothesize the selection imposed by human immune system might have played an important role here given the significant fitness cost of high-risk HPVs to the human host.

This leads to the next question, what genomic changes in evolution confer such high pathogenicity during PVs' association with human. In our analysis, we found many SNVs sites in the HPV genome showing strong segregating pattern in terms of their genotypes between low-risk and high-risk/probably high-risk HPVs, which further agreed well with the phylogenetic bifurcation between these two groups of HPVs. We consider these SNVs as risk-defining SNVs and argue that they are highly informative in telling high-risk and low-risk HPVs apart even at the single site level, without knowing any information from the rest of the HPV genome. These sites have important application value in both primary (vaccination) and secondary (screening) prevention against high-risk HPVs. As a proof of concept, we brought in four independent validation cohorts containing >6300 HPV genomes in total and found our identified risk-defining SNVs well-performed in differentiate low-risk versus high-risk HPVs in all four cohorts. Mechanistically, we found both mutational and selective forces likely contributed to the segregating genotype pattern at these sites. For example, we detected physical association between these sites and host-integration breakpoints of high-risk HPVs, suggesting microhomology-mediated DNA repair as a potential mutagenic source in shaping this pattern. Meanwhile, we found some of these risk-defining SNV sites are under significant diversifying selection, reinforcing the genotype segregation between high-risk and low-risk HPVs. As for specific selection scheme, we found the specific genotypes carried by high-risk HPVs can contribute to a higher potential of immune evasion, which explains why they are

associated with high-risk pathogenicity. Collectively, these findings highlight how the virus-host interaction is shaping HPVs' genome and pathogenicity evolution.

Given the strong association between high-risk HPV infection and malicious human cancers such as cervical and oropharyngeal cancers, discovery and application of clinically relevant HPV genomic variants in high-risk HPVs could have great value in the prevention and treatment of HPV-associated cancers. In this study, we provided proof-of-concept examples demonstrating how these variants can be used in predicting HPVs' pathogenic risks as well as the prognosis of HPV-associated cancers. While the clinical cohorts collected in our current study are still relatively limited, it is nevertheless encouraging to see how some of our identified key HPV variants can be informative in real clinical settings. We expect future studies with larger and broader HPV-associated clinical cohorts from more diverse ethnical groups can further enrich our understanding on the biological and clinical significance of these key HPV variants in the age of precision medicine.

Finally, most existing PV studies, including our current one, are anthropocentric, meaning that our sampling and knowledge of PVs are heavily biased towards human. In comparison, our knowledge of the genetic and ecological diversity of animal PVs are still rather limited, which prevents a high-resolution view into their deep evolutionary history. Recently, several large national and international consortia have been formed to sequence and characterize genomes of a wide variety of eukaryotic life on earth[40,41]. Aside from contemporary animals, researchers have also began to look for trace of PV infection in ancient animals [42]. Taken together, it can be expected that our knowledge on the virome of the life on earth could be extensively expanded in both taxonomically and historical sampling. When such data becomes available, the community will be able to assemble a more balanced and complete view on PVs regarding its biology and evolution.

Materials and Methods

PV genome collection and curation for publicly available data

A total of 2969 complete HPV genomes (quality control criteria: assembly gap <12.5% of the total length) were downloaded from the Genbank database hosted by the National Center for Biotechnology Information (NCBI) via the NCBI EFetch utility[43] (v16.2) with the query term of "human papillomavirus complete genome" (last accessed by 2023.06.01). Another 360 HPVs genomes (quality control criteria: assembly gap <12.5% of the total length) were retrieved by contacting the authors of a recently published study[34]. In addition, 233 animal PVs (including 30 non-human primate PVs, 172 non-primate mammal PVs, 6 reptile PVs, 11 bird PVs, and 14 fish PVs) were retrieved from the PaVE database[18,36] (last accessed on 2023.06.01). The metadata regarding PVs' host species origin, PVs' anatomic body parts and tissue origin, patients' geography origin, PV's taxonomic classification, PV's risk classification, patients' disease type classification, etc. was further retrieved and curated by referencing back to the original studies that sequenced these PV genomes. Given the different starting site choice and annotation quality of these PV genomes, we further reset the starting site of each assembly as the first base after the L1 stop codon and performed gene annotation with PuMA[44] (v1.2.2). The software gffread[45] (v0.9.12) was further used to extract the coding sequences (CDSs) from each annotated gene accordingly.

3548-PV-genome-based whole-genome alignment and phylogenetic tree building

All curated PV genomes were processed with windowmasker[46] (v1.0.0) for soft-masking. MAFFT[47] (v7.310; options: --auto, --preservecase) was used to generate a preliminary whole-genome alignment for the soft-masked 3548 curated human and animal PV genomes (including 2968 HPVs, 30 non-human primate PVs, 172 non-primate mammal PVs, 6 reptile PVs, 11 bird PVs). The 14 fish PV genomes were excluded due to their substantially large sequence divergence with other PVs, which tends to compromise the alignment quality. The resulting alignment was processed by ClipKIT[48] (v1.3.0; options: -m gappy

-g 0.9) before being fed to IQ-TREE[49] (v2.0.6; options: --seed 20230417 -m MFP -B 1000 --bnni -alrt 1000 --safe) to generate the guide tree, which was further used to generate the second-round whole-genome alignment with Cactus[50] (v2.2.0) by using the start-reset HPV16 (hom_sap_HPV16; NC_001526.4 with start-reset) as the reference genome. The Cactus derived hal-formatted alignment were converted to the maf format by the hal2maf tool from the HAL package[51] (v2.2.0; options: --onlyOrthologs --noDupes --maxRefGap 2000 --refGenome hom_sap_HPV16), which was further converted to the fasta format. The resulting fasta-formatted genome alignment was clipped by ClipKIT[48] (v1.3.0; option: -m gappy) and fed to IQ-TREE[49] (v2.0.6; options: --seed 20230421 -m MFP -B 1000 --bnni -alrt 1000 --safe) for the final tree building. The resulting tree was rerooted by the nw_reroot tool from the Newick Utilities[52] (v1.6.0) with PVs from birds and reptiles as outgroups. The final phylogenetic tree was visualized using FigTree (v1.4.4; http://tree.bio.ed.ac.uk/software/figtree/) and ggtree[53] (v3.6.2).

600-PV-genome-based whole-genome alignment, phylogenetic tree building, and molecular dating analyses

Based on the phylogenetic structure and host species distribution of 3548-PV-genome-based phylogenetic tree, we further generated a down-sampled dataset with 600 representative PV genomes. Such PV genome down-sampling was performed within each type-specific HPV clade for retaining the overall phylogenetic characteristics while reducing the computational burden for downstream analysis. Whole-genome alignment and tree building were further carried out based on this down-sampled 600 PV genomes following the same procedures that we used for the 3548-PV-genome-based analysis. According to previously estimated PV evolutionary rates[54], we used a Bayesian Markov Chain Monte Carlo (MCMC) based framework implemented by BEAST2[55] (v2.5.2) for molecular dating analysis. Following the suggestions from the similar analyses that have done previously[12,56,57], we used a combination of Yule tree prior and relaxed uncorrelated lognormal distribution (UCLD) of the molecular clock model for our BEAST analysis. Regarding the calibration nodes, we introduced four calibration point along the

phylogenetic tree based on a collective consideration of the general codivergence histories between PVs and their hosts, as well as the PV-specific the early pre-divergence of the Alpha, Beta, and Gamma superclades in non-primate mammal hosts: (1) the separation between the reptile PVs (represented by hem_fre_PV1/2) and Aves PVs (represented by che_myd_PV1) at 280.0 (274.9–286.8) MYA, (2) the split between reptile/bird PVs (e.g., hem_fre_PV1/2, lar_smi_PV1, pyg_ade_PV1/2) and mammalian PVs (e.g., rou_aeg_PV1, equ_cab_PV1) at 319.0 (95% CI of 316.0–322.4) MYA, (3) the split between new-world monkey PVs (e.g., sai_sci_PV1/2/3, alo_gua_PV1) and old-world monkey PVs (e.g., mac_mul_PV3, mac_fus_PV2) at 43.0 (95% CI of 40.0–44.2) MYA, and (4) the split between the lineage-specific human (e.g., hom_sap_HPV13) and chimpanzee PV (e.g., pan_pan_PV1) at 6.2 (6.0–6.5) MYA. These divergence time were retrieved from the TimeTree database[58] (v5; https://timetree.org/). Two independent MCMC runs were launched for 200000000 MCMC steps, with 10% burn-in and a subsampling frequency of 20000 generations. A minimal effective sample size (ESS) of 200 was used for evaluating the MCMC chain convergence, which were inspected in Tracer (v1.7.1; http://tree.bio.ed.ac.uk/software/tracer/). TreeAnnotator (part of with BEAST2 package) was used to generate the consensus tree with 10% burn-in. The resulting consensus tree was visualized with FigTree (v1.4.4; http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenetic generalized least squares (PGLS) analysis

The comparative data points are not expected to be independent because of their shared phylogenetic history[59], therefore we estimated the correlation of the GC% between PVs and their hosts used Phylogenetic Generalized Least Squares (PGLS) regression[60] with the *gls* function implemented in the *R* package nlme[61]. The covariance (correlation) structure was fitted using the Brownian evolution model: correlation = corBrownian(1, phy = PV_tree, form = ~PV_Name).

Codon usage analysis

To detect the codon usage bias between PVs and their hosts, we downloaded the host coding sequences from the Ensembl Metazoa database[62] (https://ftp.ensemblgenomes.ebi.ac.uk/pub/metazoa/release-57/fasta/). For PVs, we used gffread[45] (v0.9.12) to extract the coding sequences. The codon usage (per 1000 codons) was then computed accordingly for the coding sequences of each PV and their host genomes using the Python script developed in a previous study[63] (https://github.com/rhondene/Codon-Usage-in-Python). For the case where there are multiple PV isolates correspond to the same host, the arithmetic mean of their per-1000 codon usages was further summarized and used for comparison. Meanwhile, we also calculated the codon adaptation index (CAI) to evaluate the co-evolution scenario between HPVs and the human host using a similar approach described previously[64]. Briefly, we first summarized the median gene expression transcripts per million (TPM) profiles of 68 human tissues from the Genotype–Tissue Expression (GTEx) database[65] (v10; https://www.gtexportal.org/). The median expression value across all tissues for each gene and the 5% with the highest expressed genes were selected as the reference set. The CDSs of each HPV gene as well as the CDSs of the reference gene sets were concatenated respectively to calculate the CAI value[66].

Variant calling and annotation

Single nucleotide variants (SNVs) calling was identified from the final alignment of 3548 PV genomes using SNP-sites[67] (v2.5.1; option: -mv). The raw SNVs set were filtered using the following criteria: (1) variants with a missing rate of >80%, (2) variants with a missing rate of >10% and a minor allele frequency (MAF) of <0.01, with the latter filtering step carried out by PLINK[68] (v1.9; options: --geno 0.1 --maf 0.01 --allow-no-sex --make-bed --const-fid). We defined the SNVs passing the first filter as the "base SNV set" and those passing both filters as the "core SNV set". Next we used Ensembl VEP (v109.3) to annotate and predict the functional impacts of these SNVs based on a minorly edited HPV16 genome annotation GFF file, in which 1) genomic features labeled as "polyA_signal_sequence", "TATA_box"

and “URR” were relabeled as ncRNA_genes (so that variants fall in these region will all be denoted as “non_coding_transcript_exon_variant” by Ensembl VEP) and 2) genes only with “CDS” annotations but not the associated “mRNA” and “exon” annotations were fixed. After the Ensembl VEP prediction, we curated the results with the following criteria: 1) For variants annotated as “non_coding_transcript_exon_variant”, we reclassified them as “regulatory_variant”; 2) For variants annotated as “upstream_gene_variant” and “downstream_gene_variant”, only the annotations correspond to the closest gene were considered; 3) For variants located between coding genes, we labeled them as “intergenic_variant”; 4) For variants associated with multiple genes (e.g., overlapping genes), we denoted the variant based on the gene with higher functional impact.

Principal component analysis

We performed principal component analysis (PCA) on the “core SNV set” using the smartpca function from the EIGENSOFT package[69] (v8.0.0). First, the VCF-format file was converted to PLINK-format files using PLINK[68] (v1.9; options: --make-bed --allow-no-sex --double-id). The generated PLINK files was then converted to input files of smartpca using the convert function from EIGENSOFT (options: outputformat: EIGENSTRAT). Finally, the resulting files were used to perform PCA analysis using smartpca (options: lsqproject: YES numoutlieriter: 0).

PV tree vs. host tree reconciliation

To infer the horizontal host switch events based on the associations between PVs and their hosts, we used Jane (v4) [70] to compare the 600-PV-genome-based tree against the time-resolved phylogenetic tree of their hosts (retrieved from the TimeTree database[58]). The host and the 600-PV trees were converted into a Jane-compatible timed tree in nexus format using the Python script available on GitHub[71] (https://github.com/andersonbrito/openData/tree/master/brito_2020_reconciliation). For the Jane analysis, a 5-million-year “time zone” was used to order different evolutionary events. Also, a set of pre-defined

relative cost parameters for different evolutionary schemes need to be provided, for which we tried multiple combinations of relative cost values (e.g., co-speciation [0], duplication [0-3], loss [0-3], and horizontal host switch [0-3]) to have a good sampling for the parameter space. In this way, a total of 64 solutions for cophylogenetic reconstruction were obtained. The host switch event that occurred in ≥80% of the solutions was defined as a high-confidence event.

Lineage- and risk-defining variant enrichment analysis

We performed the lineage- and risk-defining variant enrichment analysis using hypergeometric test on the "base SNV set" defined previously with the Benjamini-Hochberg false discovery rate (FDR) correction. For lineage-defining variants, we identified them using the following criteria: 1) FDR-corrected *P*-value <0.01 in the focal vs. background hypergeometric test, 2) variant allele frequency ≥90% in the focal lineage, 3) variant allele frequency <20% in all non-focal lineages. Note that because the HPV16 genome that we used as the reference belongs to the Alpha PVs, the lineage-defining variants for Alpha PVs were identified as those got progressively depleted in this clade. For risk-defining variants, we scanned for variants that are enriched in either low-risk HPV groups or high-risk HPV groups. For the former, we used the criteria of 1) FDR-corrected *P*-value <0.01 in the low-risk HPV group vs. probably high-risk/high-risk HPV group hypergeometric test, 2) variant allele frequency ≥95% in the low-risk HPV group, 3) variant allele frequency <10% in the high-risk HPV group. For the latter, we used the criteria: the criteria of 1) FDR-corrected *P*-value <0.01 in the high-risk and probably high-risk HPV group vs. low-risk HPV group hypergeometric test, 2) variant allele frequency ≥95% in the high-risk and probably high-risk HPV group, 3) variant allele frequency <10% in the low-risk HPV group. Additionally, we further require the major allele genotype frequency in the high-risk and probably high-risk HPV group to be >=80%. Finally, for the case where multiple SNVs hitting the same codon, we used get_MNV (v1.0.0; https://github.com/PathoGenOmics-Lab/get_MNV) to identify and relabel the corresponding amino acid changes accordingly.

Branch- and site-specific selection analysis

We used aBSREL[25] from the HyPhy package[72] (v2.5.60) to detect episodic diversifying selection in a branch-specific manner along the 3548-PV-genome-based phylogenetic tree and its associated alignment for the Alpha-super-clade and its subtrees, with the following options: --kill-zero-lengths No --branches Foreground. After the inference, the mean omega (ω) of the tested branch and the empirical Bayes factor (EBF) value of each codon was obtained using HyPhy Vision (http://vision.hyphy.org/). In parallel, we also used the mixed effects model of evolution (MEME) [21] implemented in HyPhy to detect specific sites underlying diversifying selection across different HPV genes (e.g., *E1*, *E2*, *E5*, *E6*, *E7*, *L1*, and *L2*).

Epitopes affinity prediction

We first retrieved information of HPV16 immune epitopes from IEDB[73] (https://www.iedb.org/), with which we further extract those peptide epitopes covering our identified high-risk SNVs for epitope binding affinity prediction. Meanwhile, we retrieved the HLA allele frequencies of nine geographical regions (Europe, North Africa, North America, South America, South Asia, Sub-Saharan Africa, East Asia, Western Asia, and Oceania) from the AFND database[74] using HLAfreq [75] (v0.0.4). For each geographical region, the top 3 alleles of each HLA gene were used for epitope binding affinity prediction. Given the previously reported strong association between MHC-II and HPV-associated cancers, we employed NetMHCIIpan[76] (v4.3) to predict MHC-II binding affinity to the epitope peptides that carrying either the wildtype or mutated form of our identified high-risk SNVs with default parameters. Wilcoxon signed-rank test (paired, two-sided) was used to compare the overall epitope-to-MHC-affinity between the wildtype and mutated forms for each identified SNV.

Protein structure prediction and visualization

To evaluate the structural impact of the mutation, the original and mutated amino acid sequences of our interested HPV16 protein were independently submitted to the AlphaFold3 server[77] (https://alphafoldserver.com/) for structure prediction. The resulting crystallographic information files (CIFs) were then imported into ChimeraX [78] (v1.8) for structural comparison and visualization. Interactions involving the mutated residue and surrounding amino acids were analyzed using the “hbonds” and “contacts” commands in ChimeraX.

Clinical sample collection

The study was approved by the ethics review board of the Sun Yat-sen University Cancer Center (SYSUCC) (Issue number: B2022-509-01). A total of 131 HPV16-positive CSCC and 63 HPV16-positive OPSCC patient samples were collected from the Tumor Biobank, Department of Gynecologic Oncology, and Department of Head and Neck Surgery of SYSUCC. All patient samples passed our quality control for DNA extraction, sequencing, and genome assembly (quality control criteria: assembly gap<12.5% of the total length). For each clinical sample, frozen tissues derived from surgically treated patients with follow-up data were included. The 8th edition American Joint Commission on Cancer (AJCC) staging manual[76] was referred to for staging the CSCC and OPSCC patients. The CSCC patients are all females, with an average age of 54 years old (ranging from 29 to 80) and most of them being clinically classified as stage I-II (89%, 116/131). The OPSCC patients are mainly male (87%, 55/63) with an average age of 55 years old (ranging from 33 to 70), most of which being staged at I-II (87%, 55/63).

Cohort validation of signature HPV SNV sites

The predictive power of the identified HPV SNV signatures for pathogenic risk was benchmarked using F1 score, precision, and recall. These metrics were calculated with a custom Python script. The classification of an allele at a given risk-defining SNV site as a "high-risk allele" was considered a

positive prediction. For this analysis, true positives (TP) were defined as the number of high-risk alleles identified in HPV strains classified as high-risk/probably high-risk types. False positives (FP) were the number of high-risk alleles found in HPV strains classified as low-risk types. False negatives (FN) were the count of non-high-risk alleles (excluding gaps) within the high-risk/probably high-risk HPVs.

HPV genome amplification and sequencing

All patient samples were tumor tissues freshly frozen in liquid nitrogen within 30 min of resection. Genomic DNA extraction was performed by using the FastPure Cell/Tissue DNA Isolation Mini Kit (Cat: DC102-01, Vazyme, China). To minimize the possibility of cross-contamination, no more than 10 samples were processed in the same batch. Based on the extracted genomic DNA, we used Touchdown PCR (TD-PCR) or Nested PCR to amplify the complete genome of HPV16 with six sets of previously published primers[77] (Supplementary Figure 19 and Supplementary Table 22) to increase the specificity and sensitivity. Between TD-PCR and Nested PCR, TD-PCR was prioritized. Our TD-PCR mix includes 25 μL Green Taq Mix (Cat: P131, Vazyme, China), 2 μL of TD-PCR forward and reverse primer (referred as F1/F2/F3_Touchdown_F/R in Supplementary Figure 19; 10 μM), ~80 ng genomic DNA and sterile DEPC-H2O (Cat: R0022, Beyotime, China), which adds up to a final volume of 50 μL. Our TD-PCR program is as follows: 95°C 3 min; touchdown phase (2 cycles: 95°C 15 sec, 65°C 15 sec, 72°C 3 min 10 sec; 2 cycles: 95°C 15 sec, 62°C 15 sec, 72°C 3 min 10 sec; 4 cycles: 95°C 15 sec, 59°C 15 sec, 72°C 3 min 10 sec); 30 cycles: 95°C 15 sec, 56°C 15 sec, 72°C 3 min 10 sec; 72°C 5 min. As for nested PCR, the first-round amplification was carried out by using 12.5 μL Green Taq Mix, 1 μL of round 1 forward and reverse primer (referred as F1/F2/F3_NestedRound1_F/R in Supplementary Figure 19; 10 μM), 40–80 ng genomic DNA and sterile DEPC-H_2O in 25 μL reaction. The PCR product of the first-round PCR was diluted with 75 μL sterile DEPC-H_2O and was used as a template for second-round amplification, which was performed in a 25-μL reaction mix containing 12.5 μL of Green Taq Mix, 1 μL of round 2 forward and reverse primer (referred as F1/F2/F3_NestedRound2_F/R; 10 μM), 5 μL template and 5 μL sterile DEPC-H_2O. The same PCR cycling program was used for both rounds: 95 °C for 5 min,

20 cycles at 95 °C for 30 sec, 56 °C for 30 sec and 72 □ for 3 min 10 sec. Final extension was set at 72 □ for 5 min. For PCR product clean-up, we used the 1× VAHTS DNA Clean Beads (Cat: N411, Vazyme, China). The obtained HPV16 amplicons, containing three purified PCR amplicon fragments which were examined by electrophoresis and further quantified by Qubit 4 Fluorometer (Thermo Fisher, USA). For each sample, we mixed the three PCR amplicon fragments with a 1:1:1 mass ratio and sent for sequencing by BGI Genomics (China) using the BGI MGISEQ-2000 platform.

HPV genome assembly

The raw sequencing reads were preprocessed using fastp[79] (v0.24.0) to remove adaptor and low-quality bases. The trimmed reads were aligned to a reference genome that contains both the host (chm13v2.0) and viral genome sequence (HPV16, NC_001526.4) using BWA[80] (v0.7.17; option: mem). Only the paired-end reads mapped to the viral genome were selected for the *de novo* genome assembly with SPAdes[81] (v4.2.0; option: --isolate). All viral reads were then scaffolded with RagTag[82] (v2.1.0). Gap filling was performed using Abyss-sealer[83] (v2.3.5, options: -b20G -k32 -k48 -k64 -k80 -k96 -k112 -k128). After gap filling, assembly polishing was performed using Pilon[84] (v1.24, options: --fix snps,indels,local). A finally quality check was performed based on the dotplot generated by mummerplot[85] (v3.5) against the HPV16 reference genome. Similarly, we reset the starting site of each assembly as the first base after the L1 stop codon and performed gene annotation with PuMA[44] (v1.2.2). For each patient-derived HPV16 genome assembly, we used MAFFT[47] (v7.310, options: --auto --preservecase) to align it to the start-reset HPV16 genome and used SNP-sites[67] (v2.5.1) to extract SNV sites, which were further annotated by Ensembl VEP[86] (v109.3).

Prognostic analysis

The prognostic analysis was carried out with the *R* package survival (v3.7.0) and further visualized with survminer (v0.5.0). The Cox proportional hazards model was used to evaluate the association between

each patient-carrying HPV16 SNV (minor allele frequency ≥1%) and overall survival. To estimate the risk of death from any cause, the only censoring event was at the time of last known follow-up, and the years since cancer diagnosis was used as the time variable. The Benjamini-Hochberg false discovery rate (FDR) was calculated based on the raw *P*-values for multiple test correction. An FDR-corrected $P \leq 0.05$ (two-tailed) was considered statistically significant. Multivariable Cox regression analysis was used to explore the association between the bearing status of our identified prognostic HPV16 variant (i.e., with or without) and overall survival, further adjusted for widely used clinical prognostic factors such as age and stage. The forestploter package[87] (v1.1.3) was used for result visualization.

Acknowledgments

This work is supported by National Natural Science Foundation of China (32470663 to Jia-Xing Yue and 32000395 to Jing Li), Guangdong Provincial Pearl River Talents Program (2019QN01Y183 to Jia-Xing Yue and 2021QN02Y168 to Jing Li), Natural Science Foundation of Guangdong Province (2022A1515010717 to Jia-Xing Yue and 2022A1515011873 to Jing Li), and Young Talents Program of Sun Yat-sen University Cancer Center (YTP-SYSUCC-0042 to Jia-Xing Yue and YTP-SYSUCC-0040 to Jing Li). The funders have not played any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Huihui Li, Jingtao Chen, and Xinxin Zhang contributed equally. Jia-Xing Yue and Jing Li conceived the study. Huihui Li, Jingtao Chen, Xinxin Zhang, Ludong Yang, Lin Feng perform the data analysis. Xinxin Zhang, Jingtao Chen, Yan Shuai, Congying Chen performed the molecular experiment. Jihong Liu, Ming Song, Xin Zhang, Yi Ding contributed to the clinical sample collection. Jia-Xing Yue and Jing Li wrote the paper. All authors reviewed and contributed to the final version of the paper.

Correspondence to JL or JXY.

Data availability statement

The curated genome assemblies, genome annotations, alignments, SNVs, raw and time-resolved phylogenetic trees of the 3562 publicly available PVs, as well as their associated metadata and bioinformatics scripts, altogether packed as Supporting Dataset 01-07, have been deposited at zenodo under the DOI of 10.5281/zenodo.10049935. The sequencing reads of the HPV16 genomes isolated from the 131 CSCC and 63 OPSCC patients collected in this study have been deposited at the Genome Sequence Archive (GSA) (https://ngdc.cncb.ac.cn/gsa/) in National Genomics Data Center[88] under the accession number PRJCA020850 and the reads accession numbers CRA013231 and CRA027844. The corresponding genome assemblies of these HPVs have been submitted to GenBase[89] (https://ngdc.cncb.ac.cn/genbase/) under the accession numbers C_AA111526.1 to C_AA111719.1. respectively.

Competing Interests

A patent application related to the findings reported in this manuscript has been filed by the authors (Application No. 202510142620.7, China, 2025.02.08). A corresponding international application under the PCT (PCT/CN2025/076515) has also been filed.

Figure legends

Figure 1. The global sampling and genomic diversity of the 3562 PV genomes. (A) The host and genus distribution of the 3562 PV genomes. (B) The geographical, disease type, and risk classification of the 3329 HPV genomes. (C) The genome size and GC% distribution of the 3562 PVs with regard to the host taxonomy groups. (D) The positive PV-host correlation in GC%. (E-H) The principal component analysis of the global PV genome diversity with regard to the PV genus classification (E), anatomic body site origin (F), anatomic tissue site origin (G), and high-risk vs. low-risk pathogenicity classification (H).

Figure 2. Evolutionary time frame of PV genome divergence and evolution. Calibrated time-resolved evolutionary tree for 600 representative PV genomes, with the branch lengths proportional to divergence time in million years ago (MYA). Key divergence time points regarding the early divergence of PVs were denoted in red, with the red dash line further highlighted concordant host shifts from non-primate mammals to primates for PVs from the Alpha, Beta, and Gamma super clades. The shades in each node represent 95% highest posterior density (HPD) interval. The time tree of fish PVs was further annexed by referencing to their host divergence time (denoted with asterisk). The tracks at the bottom show the phylogenetic distribution of PVs regarding their host species group, genus classification, tissue origin, and risk classification.

Figure 3. Phylogenetically reconciled evolutionary events during PV evolution. (A) The number of different types of evolutionary events inferred under a range of event-specific cost parameter settings. CO: cospeciation, HS: host switch, DUP: duplication; LO: loss. (B) Horizontal host switch events revealed by reconciliation between the virus and host phylogenetic trees. The phylogenetic node IDs associated with these host switch events were indicated, which were referred to in Supplementary Table 6. Geologically defined time zones are further depicted underneath. C: Carboniferous; P: Permian; Tr: Triassic; J: Jurassic; K: Cretaceous; Pg: Paleogene; Ng: Neogene. Key divergence time point such as 96.7 MYA for Alpha vs. Beta/Gamma PVs and 69.5 MYA for non-primate to primate host switch were also indicated.

Figure 4. Key mutations defining the early divergence of the Alpha-, Beta-, and Gamma-PVs. (A) The prevalence heatmap of identified key mutations got progressively enriched in the Beta-and-Gamma-super-clades (▯), Beta- (▯), and Gamma-super-clades (▯), or progressively depleted in the Alpha-super-clades (▯). The genome and protein coordinates of these mutations were listed on the right, with the red fonts indicating mutations under significant diversifying selection (*P<0.1, **P<0.05, ***P<0.01). (B) The phylogenetic distribution and mutational directions of key mutations under diversifying selection that drove the super-clade-specific diversification.

Figure 5. Risk-defining signature SNV sites with contrasting genotypes between low-risk and high-risk HPVs. (A) The phylogenetic tree of 3548 PVs. (B) The metadata track for the 3548 PVs regarding their species origin, pathogenic risk, and associated diseases. (C) The nucleotide state of risk-defining signature SNV sites in *E1* and *E2*. Sites highlighted in red are subject to significant diversifying selection evaluated by HyPhy MEME. (D) The physical distribution of risk-defining SNV sites and the breaking points of the HPV genome. (E-G) The MHC-II binding affinity comparison of epitopes between high-risk and low-risk epitopes in E1 (E, F) and E2 (G). The amino acid changes in epitope peptides (underlined sequences) were indicated above, while the MHC-II binding affinity [measured by 1/(%rank_score)] for the top 3 MHC-II alleles from nine geographical regions were plotted below. The MHC-II allele with the highest binding affinity were highlighted. The *P*-value of Wilcoxon signed-rank test (paired, two-sided) is given. Dashed lines represent the thresholds for strong binding (SB, 1%) and weak binding (WB, 5%).

Figure 6. Performance validation of signature HPV SNV sites for pathogenic risk prediction. (A) The prediction performance of each individual risk-associated signature site in identifying HPVs with high-risk or probably high-risk pathogenicity in cohorts retrieved from Genbank (left: cohort 1; right: cohort 2). (B) The prediction performance of each individual risk-associated signature site in identifying HPVs with high-risk or probably high-risk pathogenicity in cohorts collected in this study (left: cohort 3; right: cohort 4). The five selection-drive sites are further highlighted in red color as well as with the asterisk notation. The genomic coordinates indicated are based on the start-reset-version of the HPV16 reference genome.

Figure 7. Discovery of prognosis-associated HPV16 variants. (A-D) The overall survival (OS) of patients stratified by individual prognostic HPV16 variants identified in our CSCC (A–C) and OPSCC (D) cohorts. (E) Prognostic factors for overall survival (OS) by multivariate analysis. (F-G) The structural view of the two nonsynonymous prognostic variants identified in *L2* (F) and *E2* (G) and their impacts on intramolecular interactions.

References

1. McBride AA. Human papillomaviruses: diversity, infection and host interactions. Nat Rev Microbiol. 2022;20: 95–108. doi:10.1038/s41579-021-00617-5

2. Van Doorslaer K, Chen Z, Bernard H-U, Chan PKS, DeSalle R, Dillner J, et al. ICTV Virus Taxonomy Profile: Papillomaviridae. Journal of General Virology. 2018;99: 989–990. doi:10.1099/jgv.0.001105

3. DiMaio D, Petti LM. The E5 proteins. Virology. 2013;445: 99–114. doi:10.1016/j.virol.2013.05.006

4. Vande Pol SB, Klingelhutz AJ. Papillomavirus E6 oncoproteins. Virology. 2013;445: 115–137. doi:10.1016/j.virol.2013.04.026

5. Roman A, Munger K. The papillomavirus E7 proteins. Virology. 2013;445: 138–168. doi:10.1016/j.virol.2013.04.013

6. de Villiers E-M, Fauquet C, Broker TR, Bernard H-U, zur Hausen H. Classification of papillomaviruses. Virology. 2004;324: 17–27. doi:10.1016/j.virol.2004.03.033

7. Bernard H-U, Burk RD, Chen Z, van Doorslaer K, Hausen H zur, de Villiers E-M. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology. 2010;401: 70–79. doi:10.1016/j.virol.2010.02.002

8. zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer. 2002;2: 342–350. doi:10.1038/nrc798

9. Viens LJ. Human Papillomavirus–Associated Cancers — United States, 2008–2012. MMWR Morb Mortal Wkly Rep. 2016;65. doi:10.15585/mmwr.mm6526a1

10. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet. 2002;32: 569–577. doi:10.1038/ng1202-569

11. Hafner MS, Nadler SA. Phylogenetic trees support the coevolution of parasites and their hosts. Nature. 1988;332: 258–259. doi:10.1038/332258a0

12. Chen Z, DeSalle R, Schiffman M, Herrero R, Wood CE, Ruiz JC, et al. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. PLOS Pathogens. 2018;14: e1007352. doi:10.1371/journal.ppat.1007352

13. Gottschling M, Göker M, Stamatakis A, Bininda-Emonds ORP, Nindl I, Bravo IG. Quantifying the phylodynamic forces driving papillomavirus evolution. Mol Biol Evol. 2011;28: 2101–2113. doi:10.1093/molbev/msr030

14. Bernard H-U, Chan S-Y, Delius H. Evolution of Papillomaviruses. In: zur Hausen H, editor. Human Pathogenic Papillomaviruses. Berlin, Heidelberg: Springer; 1994. pp. 33–54. doi:10.1007/978-3-642-78487-3_3

15. Van Doorslaer K. Evolution of the Papillomaviridae. Virology. 2013;445: 11–20. doi:10.1016/j.virol.2013.05.012

16. Bravo IG, Alonso Á. Phylogeny and evolution of papillomaviruses based on the E1 and E2 proteins. Virus Genes. 2007;34: 249–262. doi:10.1007/s11262-006-0017-4

17. Shah SD, Doorbar J, Goldstein RA. Analysis of Host–Parasite Incongruence in Papillomavirus Evolution Using Importance Sampling. Mol Biol Evol. 2010;27: 1301–1314. doi:10.1093/molbev/msq015

18. Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, et al. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Research. 2013;41: D571–D578. doi:10.1093/nar/gks984

19. Yuan H, Myers S, Wang J, Zhou D, Woo JA, Kallakury B, et al. Use of Reprogrammed Cells to Identify Therapy for Respiratory Papillomatosis. New England Journal of Medicine. 2012;367: 1220–1227. doi:10.1056/NEJMoa1203055

20. Willemsen A, Bravo IG. Origin and evolution of papillomavirus (onco)genes and genomes. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374: 20180303. doi:10.1098/rstb.2018.0303

21. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genetics. 2012;8: e1002764. doi:10.1371/journal.pgen.1002764

22. Carter JJ, Wipf GC, Benki SF, Christensen ND, Galloway DA. Identification of a Human Papillomavirus Type 16-Specific Epitope on the C-Terminal Arm of the Major Capsid Protein L1. Journal of Virology. 2003;77: 11625–11632. doi:10.1128/jvi.77.21.11625-11632.2003

23. Day PM, Thompson CD, Buck CB, Pang Y-YS, Lowy DR, Schiller JT. Neutralization of Human Papillomavirus with Monoclonal Antibodies Reveals Different Mechanisms of Inhibition. Journal of Virology. 2007;81: 8784–8792. doi:10.1128/jvi.00552-07

24. Posth C, Wißing C, Kitagawa K, Pagani L, van Holstein L, Racimo F, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. Nat Commun. 2017;8: 16046. doi:10.1038/ncomms16046

25. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Molecular Biology and Evolution. 2015;32: 1342–1353. doi:10.1093/molbev/msv022

26. Porter VL, Ng M, O'Neill K, MacLennan S, Corbett RD, Culibrk L, et al. Rearrangements of viral and human genomes at human papillomavirus integration events and their allele-specific impacts on cancer genome regulation. Genome Res. 2025;35: 653–670. doi:10.1101/gr.279041.124

27. Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat Genet. 2015;47: 158–163. doi:10.1038/ng.3178

28. Chen D, Juko-Pecirep I, Hammer J, Ivansson E, Enroth S, Gustavsson I, et al. Genome-wide association study of susceptibility loci for cervical cancer. J Natl Cancer Inst. 2013;105: 624–633. doi:10.1093/jnci/djt051

29. Shi Y, Li L, Hu Z, Li S, Wang S, Liu J, et al. A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. Nat Genet. 2013;45: 918–922. doi:10.1038/ng.2687

30. Lesseur C, Diergaarde B, Olshan AF, Wünsch-Filho V, Ness AR, Liu G, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. Nat Genet. 2016;48: 1544–1550. doi:10.1038/ng.3685

31. Lei J, Ploner A, Lagheden C, Eklund C, Kleppe SN, Andrae B, et al. High-risk human papillomavirus status and prognosis in invasive cervical cancer: A nationwide cohort study. PLOS Medicine. 2018;15: e1002666. doi:10.1371/journal.pmed.1002666

32. Lei J, Arroyo-Mühr LS, Lagheden C, Eklund C, Nordqvist Kleppe S, Elfström M, et al. Human Papillomavirus Infection Determines Prognosis in Cervical Cancer. JCO. 2022;40: 1522–1528. doi:10.1200/JCO.21.01930

33. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. N Engl J Med. 2010;363: 24–35. doi:10.1056/NEJMoa0912217

34. Kuhs KAL, Faden DL, Chen L, Smith DK, Pinheiro M, Wood CB, et al. Genetic variation within the human papillomavirus type 16 genome is associated with oropharyngeal cancer prognosis. Annals of Oncology. 2022;33: 638–648. doi:10.1016/j.annonc.2022.03.005

35. Doorbar J, Egawa N, Griffin H, Kranjec C, Murakami I. Human papillomavirus molecular biology and disease association. Reviews in Medical Virology. 2015;25: 2–23. doi:10.1002/rmv.1822

36. Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Research. 2017;45: D499–D506. doi:10.1093/nar/gkw879

37. Lehoux M, Fradet-Turcotte A, Lussier-Price M, Omichinski JG, Archambault J. Inhibition of Human Papillomavirus DNA Replication by an E1-Derived p80/UAF1-Binding Peptide. Journal of Virology. 2012;86: 3486–3500. doi:10.1128/jvi.07003-11

38. Côté-Martin A, Moody C, Fradet-Turcotte A, D'Abramo CM, Lehoux M, Joubert S, et al. Human Papillomavirus E1 Helicase Interacts with the WD Repeat Protein p80 To Promote Maintenance of the Viral Genome in Keratinocytes. Journal of Virology. 2008;82: 1271–1283. doi:10.1128/jvi.01405-07

39. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12: 32–42. doi:10.1038/nrg2899

40. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences. 2018;115: 4325–4333. doi:10.1073/pnas.1720115115

41. The Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proceedings of the National Academy of Sciences. 2022;119: e2115642118. doi:10.1073/pnas.2115642118

42. Larsen BB, Cole KL, Worobey M. Ancient DNA provides evidence of 27,000-year-old papillomavirus infection and long-term codivergence with rodents. Virus Evolution. 2018;4: vey014. doi:10.1093/ve/vey014

43. Kans J. Entrez Direct: E-utilities on the Unix Command Line. Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US); 2023. Available: https://www.ncbi.nlm.nih.gov/books/NBK179288/

44. Pace J, Youens-Clark K, Freeman C, Hurwitz B, Van Doorslaer K. PuMA: A papillomavirus genome annotation tool. Virus Evolution. 2020;6: veaa068. doi:10.1093/ve/veaa068

45. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research. 2020. p. ISCB Comm J-304. doi:10.12688/f1000research.23297.2

46. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006;22: 134–141. doi:10.1093/bioinformatics/bti774

47. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066. doi:10.1093/nar/gkf436

48. Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020;18: e3001007. doi:10.1371/journal.pbio.3001007

49. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32: 268–274. doi:10.1093/molbev/msu300

50. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 2011;21: 1512–1528. doi:10.1101/gr.123356.111

51. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics. 2013;29: 1341–1342. doi:10.1093/bioinformatics/btt128

52. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics. 2010;26: 1669–1670. doi:10.1093/bioinformatics/btq243

53. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8: 28–36. doi:10.1111/2041-210X.12628

54. Rector A, Lemey P, Tachezy R, Mostmans S, Ghim S-J, Van Doorslaer K, et al. Ancient papillomavirus-host co-speciation in Felidae. Genome Biology. 2007;8: R57. doi:10.1186/gb-2007-8-4-r57

55. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLOS Computational Biology. 2014;10: e1003537. doi:10.1371/journal.pcbi.1003537

56. Chen Z, Long T, Wong PY, Ho WCS, Burk RD, Chan PKS. Non-human Primate Papillomaviruses Share Similar Evolutionary Histories and Niche Adaptation as the Human Counterparts. Front Microbiol. 2019;10: 2093. doi:10.3389/fmicb.2019.02093

57. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol. 2012;29: 2157–2167. doi:10.1093/molbev/mss084

58. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, et al. TimeTree 5: An Expanded Resource for Species Divergence Times. Molecular Biology and Evolution. 2022;39: msac174. doi:10.1093/molbev/msac174

59. Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist. 1985;125: 1–15.

60. Smaers JB, Rohlf FJ. Testing species' deviation from allometric predictions using the phylogenetic regression. Evolution. 2016;70: 1145–1149. doi:10.1111/evo.12910

61. Pinheiro J, Bates D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. 2025. Available: https://CRAN.R-project.org/package=nlme

62. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Research. 2023;51: D933–D941. doi:10.1093/nar/gkac958

63. Wint R, Cleary MD. Transfer RNA Levels Are Tuned to Support Differentiation During Drosophila Neurogenesis. Genes. 2024;15: 1602. doi:10.3390/genes15121602

64. Wu X, Shan K, Zan F, Tang X, Qian Z, Lu J. Optimization and Deoptimization of Codons in SARS-CoV-2 and Related Implications for Vaccine Development. Advanced Science. 2023;10: 2205445. doi:10.1002/advs.202205445

65. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653

66. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987;15: 1281–1295. doi:10.1093/nar/15.3.1281

67. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2: e000056. doi:10.1099/mgen.0.000056

68. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81: 559–575.

69. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genetics. 2006;2: e190. doi:10.1371/journal.pgen.0020190

70. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol. 2010;5: 1–10. doi:10.1186/1748-7188-5-16

71. Brito AF, Baele G, Nahata KD, Grubaugh ND, Pinney JW. Intrahost speciations and host switches played an important role in the evolution of herpesviruses. Virus Evolution. 2021;7: veab025. doi:10.1093/ve/veab025

72. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Molecular Biology and Evolution. 2020;37: 295–299. doi:10.1093/molbev/msz197

73. Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Research. 2025;53: D436–D443. doi:10.1093/nar/gkae1092

74. Gonzalez-Galarza FF, McCabe A, Santos EJMD, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. 2020;48: D783–D788. doi:10.1093/nar/gkz1029

75. Wells DA, McAuley M. HLAfreq: Download and combine HLA allele frequency data. bioRxiv; 2023. p. 2023.09.15.557761. doi:10.1101/2023.09.15.557761

76. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Research. 2020;48: W449–W454. doi:10.1093/nar/gkaa379

77. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630: 493–500. doi:10.1038/s41586-024-07487-w

78. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF ChimeraX: Tools for structure building and analysis. Protein Science. 2023;32: e4792. doi:10.1002/pro.4792

79. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

80. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv; 2013. doi:10.48550/arXiv.1303.3997

81. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics. 2020;70: e102. doi:10.1002/cpbi.102

82. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biology. 2022;23: 258. doi:10.1186/s13059-022-02823-7

83. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 2017;27: 768–777. doi:10.1101/gr.214346.116

84. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014;9: e112963. doi:10.1371/journal.pone.0112963

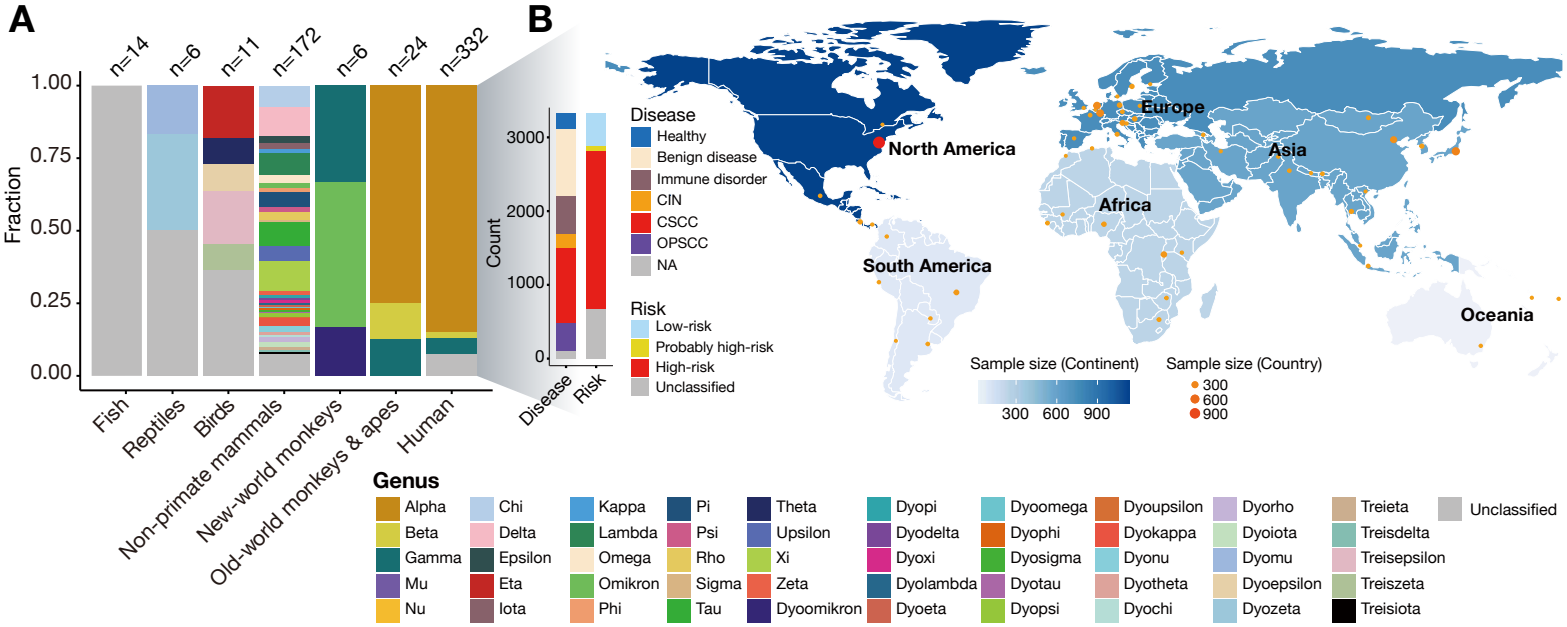
85. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018;14: e1005944. doi:10.1371/journal.pcbi.1005944

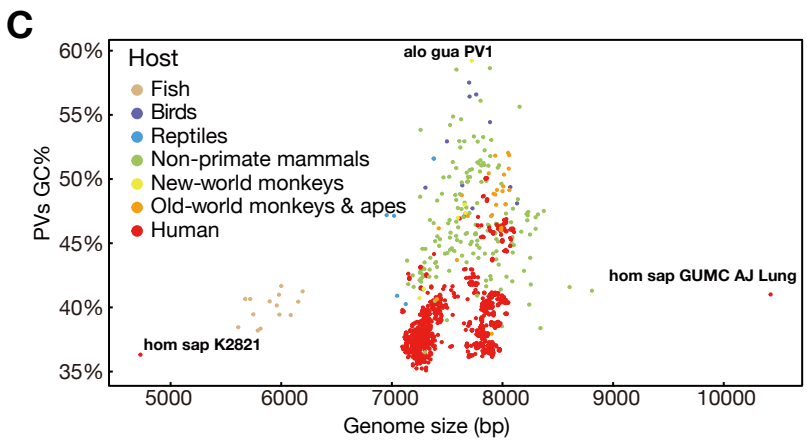
86. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17: 122. doi:10.1186/s13059-016-0974-4

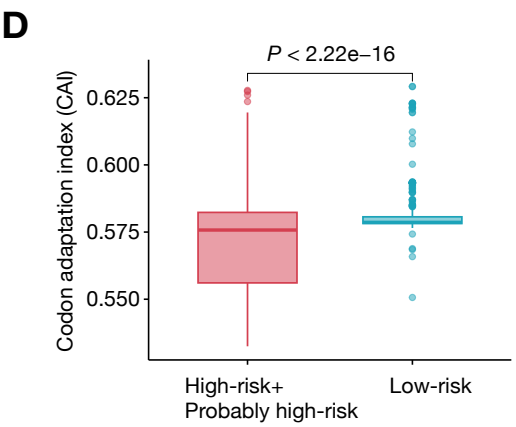
87. Dayimu A. forestploter: Create a Flexible Forest Plot. 2025. Available: https://cran.r-project.org/web/packages/forestploter/index.html

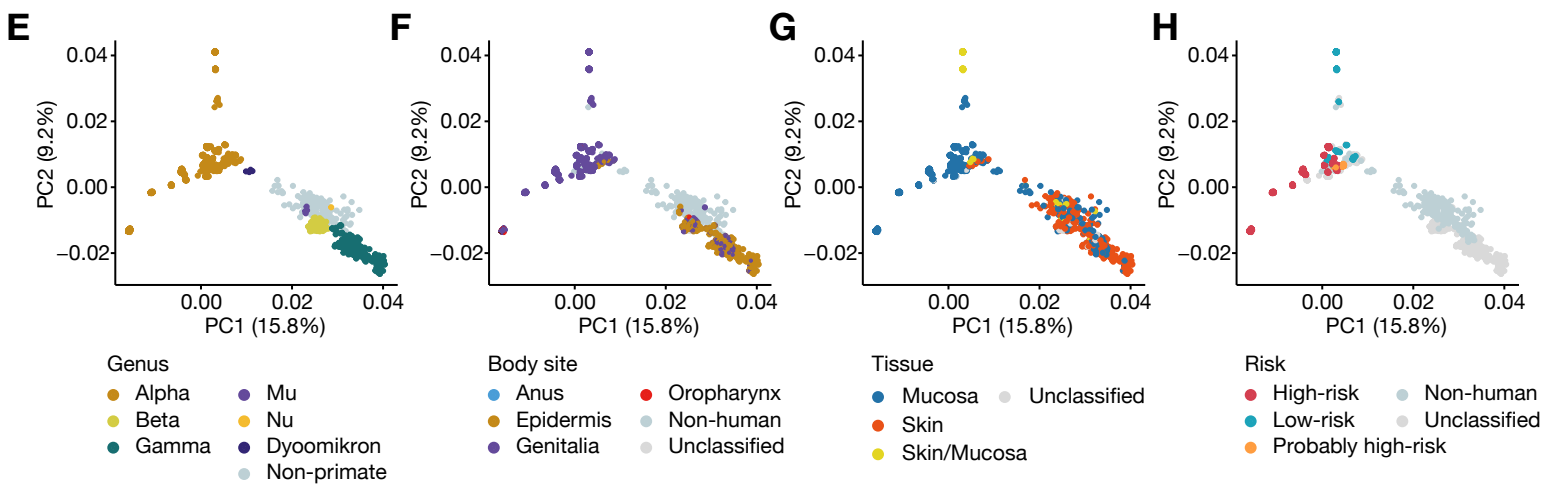
88. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Research. 2025;53: D30–D44. doi:10.1093/nar/gkae978

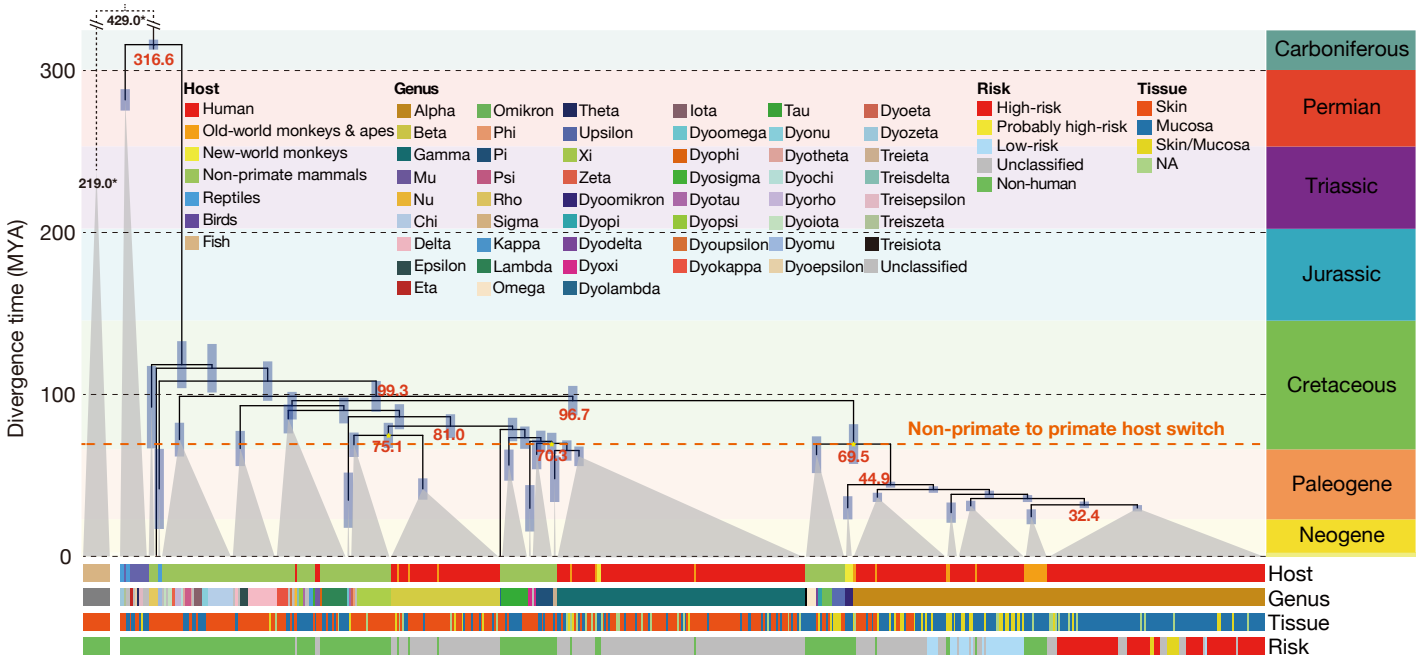
89. Bu C, Zheng X, Zhao X, Xu T, Bai X, Jia Y, et al. GenBase: A Nucleotide Sequence Database. Genomics, Proteomics & Bioinformatics. 2024;22: qzae047. doi:10.1093/gpbjnl/qzae047

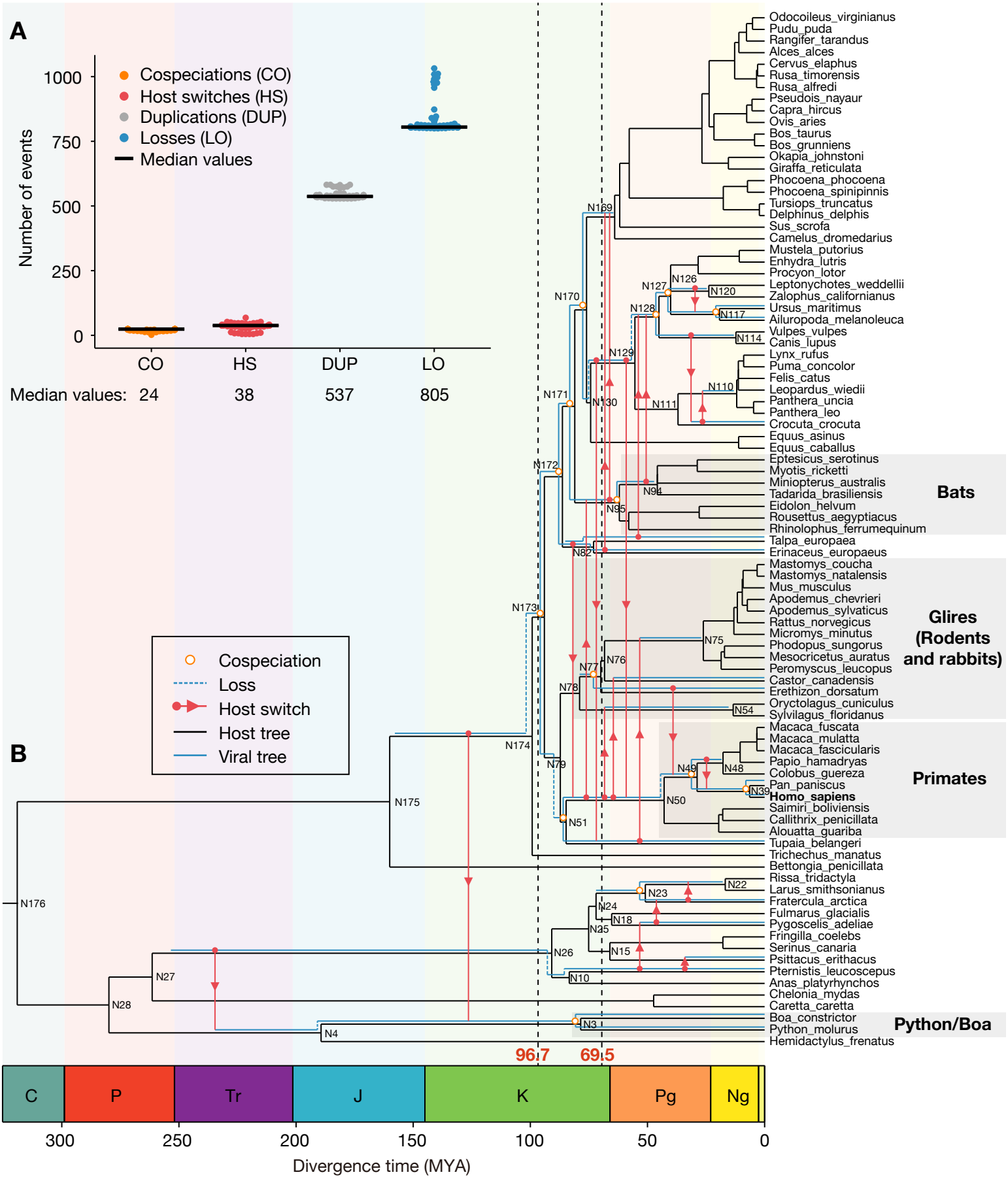

A
n=14
n=6
n=11
n=172
n=6
n=24
n=332
Fraction
1.00
0.75
0.50
0.25
0.00
Fish
Reptiles
Birds
Non-primate mammals
New-world monkeys
Old-world monkeys & apes
Human
Count
3000
2000
1000
0
Disease
Risk
B
Disease
Healthy
Benign disease
Immune disorder
CIN
CSCC
OPSCC
NA
Risk
Low-risk
Probably high-risk
High-risk
Unclassified
North America
Europe
Asia
Africa
South America
Oceania
Sample size (Continent)
300 600 900
Sample size (Country)
300
600
900
Genus
Alpha
Beta
Gamma
Mu
Nu
Chi
Delta
Epsilon
Eta
Iota
Kappa
Lambda
Omega
Omikron
Phi
Pi
Psi
Rho
Sigma
Tau
Theta
Upsilon
Xi
Zeta
Dyoomikron
Dyopi
Dyodelta
Dyoxi
Dyolambda
Dyoeta
Dyoomega
Dyophi
Dyosigma
Dyotau
Dyopsi
Dyoupsilon
Dyokappa
Dyonu
Dyotheta
Dyochi
Dyorho
Dyoiota
Dyomu
Dyoepsilon
Dyozeta
Treieta
Treisdelta
Treisepsilon
Treiszeta
Treisiota
Unclassified

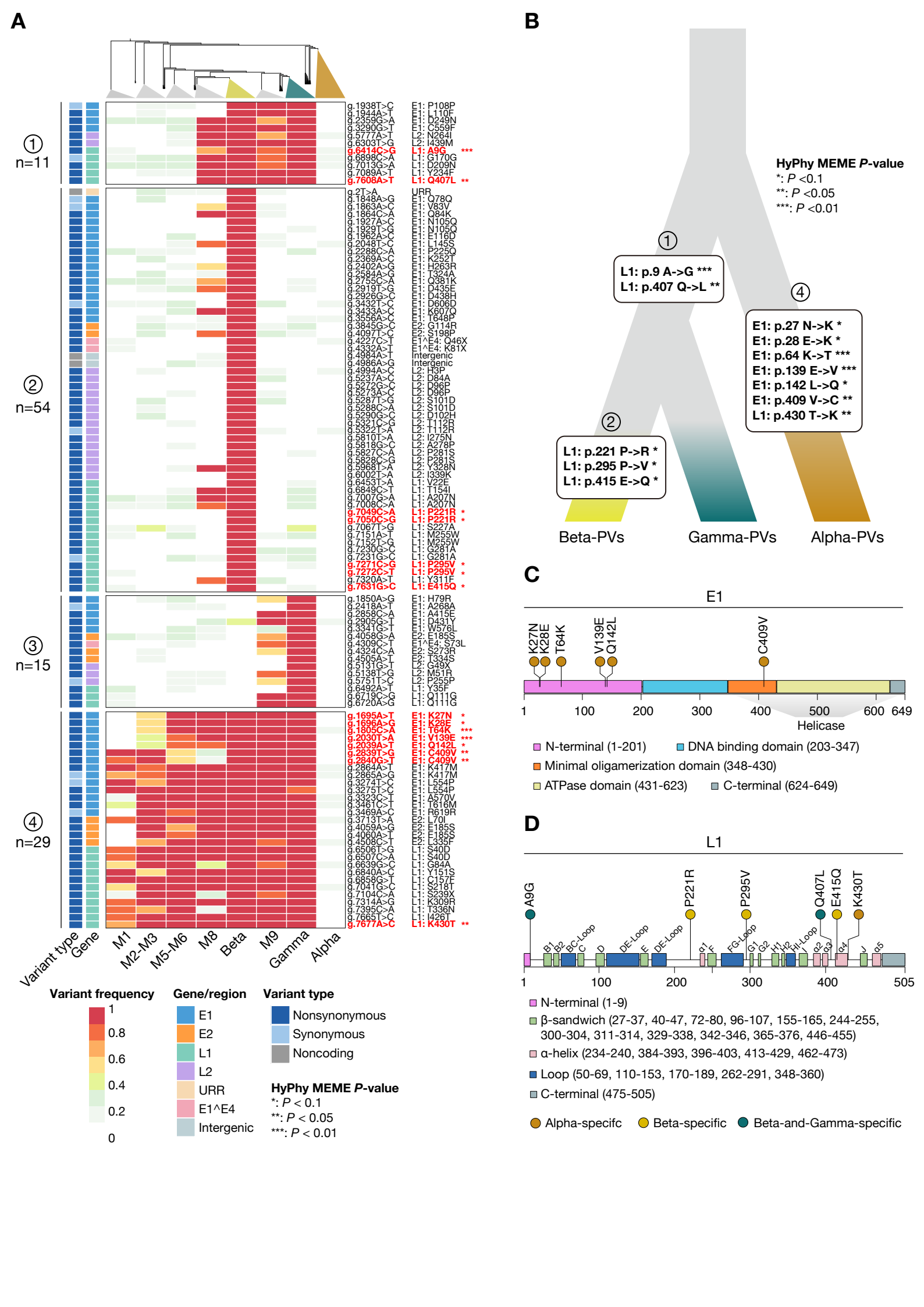

C
Host
Fish
Birds
Reptiles
Non-primate mammals
New-world monkeys
Old-world monkeys & apes
Human
alo gua PV1
hom sap GUMC AJ Lung
hom sap K2821
PVs GC%
60%
55%
50%
45%
40%
35%
5000
6000
7000
8000
9000
10000
Genome size (bp)

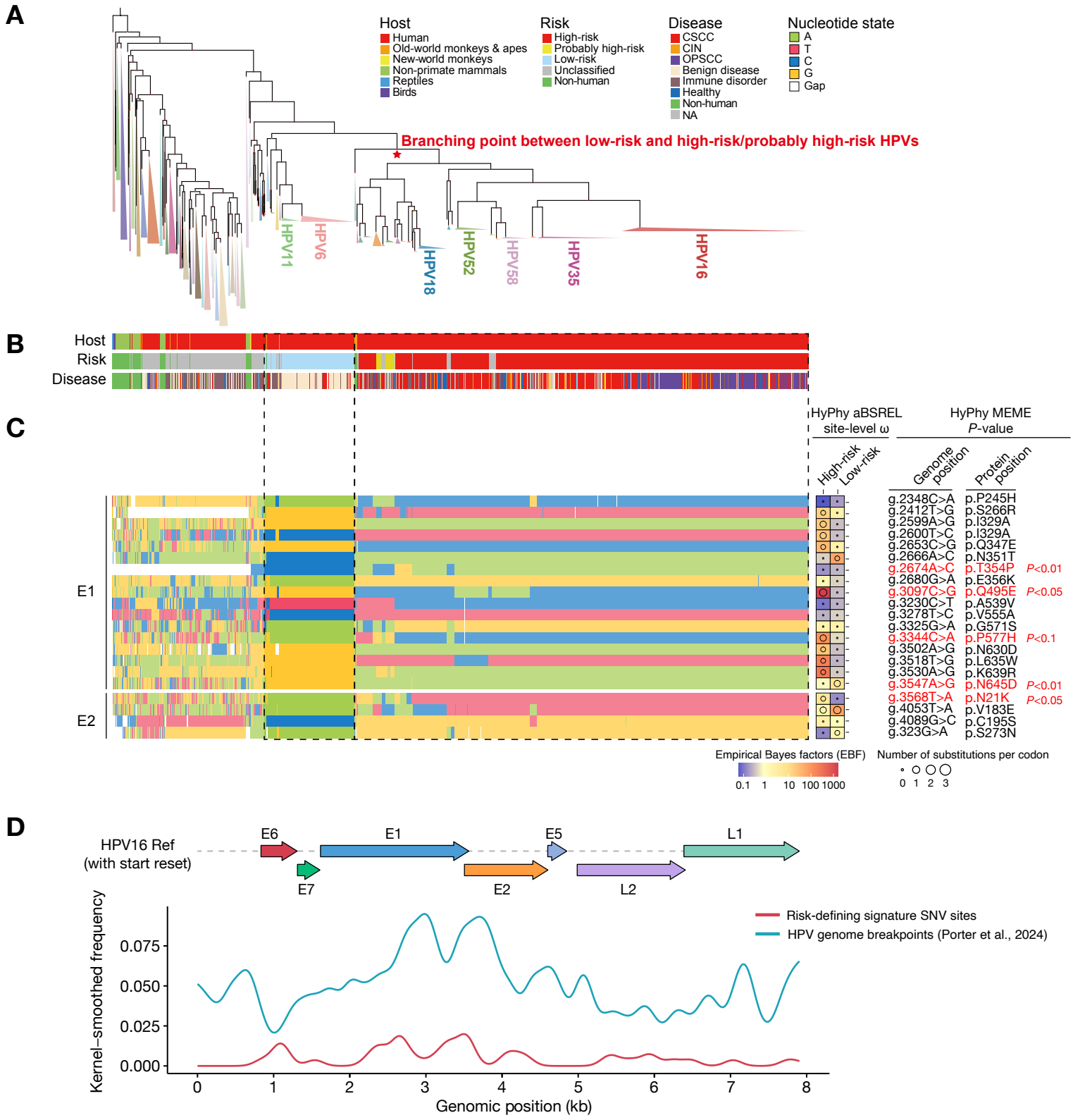

D
P < 2.22e−16
Codon adaptation index (CAI)
0.625
0.600
0.575
0.550
High-risk+
Probably high-risk
Low-risk

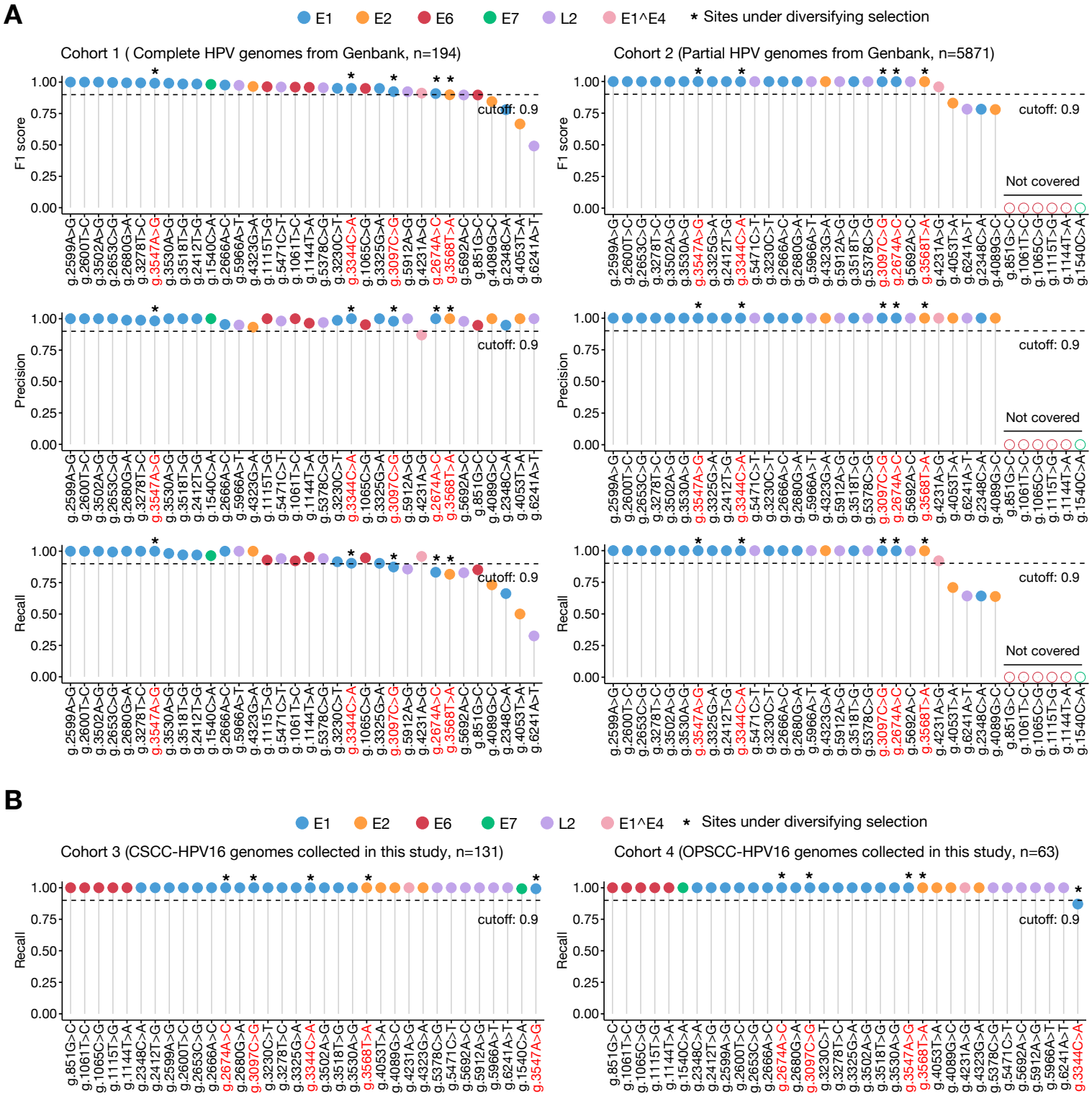

E
F
G
H
PC2 (9.2%)
PC1 (15.8%)
0.04
0.02
0.00
−0.02
0.00
0.02
0.04
Genus
Alpha
Beta
Gamma
Mu
Nu
Dyoomikron
Non-primate
Body site
Anus
Epidermis
Genitalia
Oropharynx
Non-human
Unclassified
Tissue
Mucosa
Skin
Skin/Mucosa
Unclassified
Risk
High-risk
Low-risk
Probably high-risk
Non-human
Unclassified

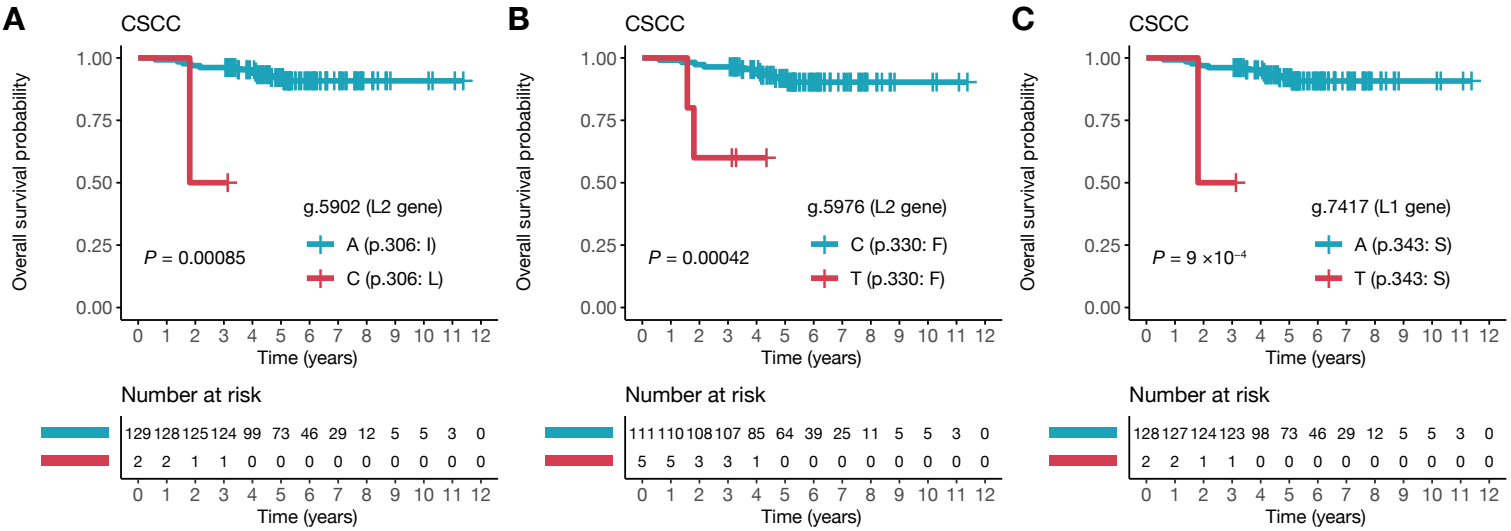

429.0*
316.6
219.0*
Divergence time (MYA)
300
200
100
0
Host
Human
Old-world monkeys & apes
New-world monkeys
Non-primate mammals
Reptiles
Birds
Fish
Genus
Alpha
Beta
Gamma
Mu
Nu
Chi
Delta
Epsilon
Eta
Omikron
Phi
Pi
Psi
Rho
Sigma
Kappa
Lambda
Omega
Theta
Upsilon
Xi
Zeta
Dyoomikron
Dyopi
Dyodelta
Dyoxi
Dyolambda
Iota
Dyoomega
Dyophi
Dyosigma
Dyotau
Dyopsi
Dyoupsilon
Dyokappa
Tau
Dyonu
Dyotheta
Dyochi
Dyorho
Dyoiota
Dyomu
Dyoepsilon
Dyoeta
Dyozeta
Treieta
Treisdelta
Treisepsilon
Treiszeta
Treisiota
Unclassified
Risk
High-risk
Probably high-risk
Low-risk
Unclassified
Non-human
Tissue
Skin
Mucosa
Skin/Mucosa
NA
99.3
96.7
81.0
75.1
70.3
69.5
44.9
32.4
Non-primate to primate host switch
Carboniferous
Permian
Triassic
Jurassic
Cretaceous
Paleogene
Neogene
Host
Genus
Tissue
Risk

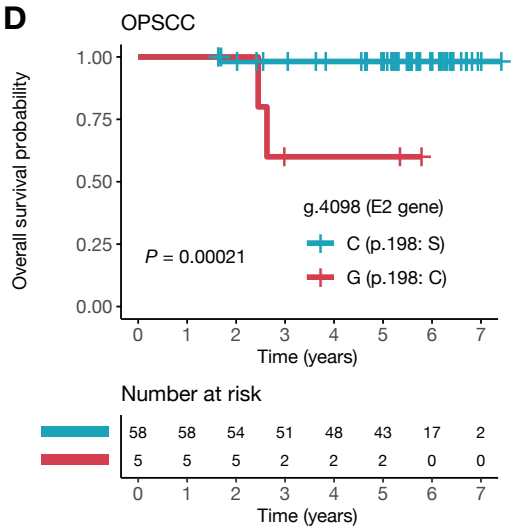

A
Cospeciations (CO)
Host switches (HS)
Duplications (DUP)
Losses (LO)
Median values
Number of events
1000
750
500
250
0
CO
HS
DUP
LO
Median values: 24 38 537 805
B
Cospeciation
Loss
Host switch
Host tree
Viral tree
Odocoileus_virginianus
Pudu_puda
Rangifer_tarandus
Alces_alces
Cervus_elaphus
Rusa_timorensis
Rusa_alfredi
Pseudois_nayaur
Capra_hircus
Ovis_aries
Bos_taurus
Bos_grunniens
Okapia_johnstoni
Giraffa_reticulata
Phocoena_phocoena
Phocoena_spinipinnis
Tursiops_truncatus
Delphinus_delphis
Sus_scrofa
Camelus_dromedarius
Mustela_putorius
Enhydra_lutris
Procyon_lotor
Leptonychotes_weddellii
Zalophus_californianus
Ursus_maritimus
Ailuropoda_melanoleuca
Vulpes_vulpes
Canis_lupus
Lynx_rufus
Puma_concolor
Felis_catus
Leopardus_wiedii
Panthera_uncia
Panthera_leo
Crocuta_crocuta
Equus_asinus
Equus_caballus
Eptesicus_serotinus
Myotis_ricketti
Miniopterus_australis
Tadarida_brasiliensis
Eidolon_helvum
Rousettus_aegyptiacus
Rhinolophus_ferrumequinum
Talpa_europaea
Erinaceus_europaeus
Mastomys_coucha
Mastomys_natalensis
Mus_musculus
Apodemus_chevrieri
Apodemus_sylvaticus
Rattus_norvegicus
Micromys_minutus
Phodopus_sungorus
Mesocricetus_auratus
Peromyscus_leucopus
Castor_canadensis
Erethizon_dorsatum
Oryctolagus_cuniculus
Sylvilagus_floridanus
Macaca_fuscata
Macaca_mulatta
Macaca_fascicularis
Papio_hamadryas
Colobus_guereza
Pan_paniscus
Homo_sapiens
Saimiri_boliviensis
Callithrix_penicillata
Alouatta_guariba
Tupaia_belangeri
Trichechus_manatus
Bettongia_penicillata
Rissa_tridactyla
Larus_smithsonianus
Fratercula_arctica
Fulmarus_glacialis
Pygoscelis_adeliae
Fringilla_coelebs
Serinus_canaria
Psittacus_erithacus
Pternistis_leucoscepus
Anas_platyrhynchos
Chelonia_mydas
Caretta_caretta
Boa_constrictor
Python_molurus
Hemidactylus_frenatus
Bats
Glires (Rodents and rabbits)
Primates
Python/Boa
N176
N175
N174
N173
N172
N171
N170
N169
N130
N129
N128
N127
N126
N120
N117
N114
N111
N110
N95
N94
N82
N78
N77
N76
N75
N54
N51
N50
N49
N48
N39
N79
N28
N27
N26
N25
N24
N23
N22
N18
N15
N10
N4
N3
96.7
69.5
C
P
Tr
J
K
Pg
Ng
300
250
200
150
100
50
0
Divergence time (MYA)


A
① n=11
② n=54
③ n=15
④ n=29
g.1938T>C E1: P108P
g.1944A>T E1: L110F
g.2359G>A E1: D249N
g.3290G>T E1: C559F
g.5777A>T L2: N264I
g.6303T>G L2: I439M
g.6414C>G L1: A9G ***
g.6898C>A L1: G170G
g.7013G>A L1: D209N
g.7089A>T L1: Y234F
g.7608A>T L1: Q407L **
g.2T>A URR
g.1848A>G E1: Q78Q
g.1863A>C E1: V83V
g.1864C>A E1: Q84K
g.1927A>C E1: N105Q
g.1929T>G E1: N105Q
g.1962A>C E1: E116D
g.2048T>C E1: L145S
g.2288C>A E1: P225Q
g.2369A>C E1: K252T
g.2402A>G E1: H263R
g.2584A>G E1: T324A
g.2755C>A E1: Q381K
g.2919T>G E1: D435E
g.2926G>C E1: D438H
g.3432T>C E1: D606D
g.3433A>C E1: K607Q
g.3556A>C E1: T648P
g.3845G>C E2: G114R
g.4097T>C E2: S198P
g.4227C>T E1^E4: Q46X
g.4332A>T E1^E4: K81X
g.4984A>T Intergenic
g.4986A>G Intergenic
g.4994A>C L2: H3P
g.5237A>C L2: D84A
g.5272G>C L2: D96P
g.5273A>C L2: D96P
g.5287T>G L2: S101D
g.5288C>A L2: S101D
g.5290G>C L2: D102H
g.5321C>G L2: T112R
g.5322T>A L2: T112R
g.5810T>A L2: I275N
g.5818G>C L2: A278P
g.5827C>A L2: P281S
g.5828C>G L2: P281S
g.5968T>A L2: Y328N
g.6002T>A L2: I339K
g.6453T>A L1: V22E
g.6849C>T L1: T154I
g.7007G>A L1: A207N
g.7008C>A L1: A207N
g.7049C>A L1: P221R *
g.7050C>G L1: P221R *
g.7067T>G L1: S227A
g.7151A>T L1: M255W
g.7152T>G L1: M255W
g.7230G>C L1: G281A
g.7231G>C L1: G281A
g.7271C>G L1: P295V *
g.7272C>T L1: P295V *
g.7320A>T L1: Y311F
g.7631G>C L1: E415Q *
g.1850A>G E1: H79R
g.2418A>T E1: A268A
g.2858C>A E1: A415E
g.2905G>T E1: D431Y
g.3341G>T E1: W576L
g.4058G>A E2: E185S
g.4309C>T E1^E4: S73L
g.4324C>A E2: S273R
g.4505A>T E2: T334S
g.5131G>T L2: G49X
g.5138T>G L2: M51R
g.5751T>C L2: P255P
g.6492A>T L1: Y35F
g.6719C>G L1: Q111G
g.6720A>G L1: Q111G
g.1695A>T E1: K27N *
g.1696A>G E1: K28E *
g.1805C>A E1: T64K ***
g.2030T>A E1: V139E ***
g.2039A>T E1: Q142L *
g.2839T>G E1: C409V **
g.2840G>T E1: C409V **
g.2864A>T E1: K417M
g.2865A>G E1: K417M
g.3274T>C E1: L554P
g.3275T>C E1: L554P
g.3323C>T E1: A570V
g.3461C>T E1: T616M
g.3469A>C E1: R619R
g.3713T>A E2: L70I
g.4059A>G E2: E185S
g.4060A>T E2: E185S
g.4508C>T E2: L335F
g.6506T>G L1: S40D
g.6507C>A L1: S40D
g.6639G>C L1: G84A
g.6840A>C L1: Y151S
g.6859G>T L1: C157F
g.7041G>C L1: S218T
g.7104C>A L1: S239X
g.7314A>G L1: K309R
g.7395C>A L1: T336N
g.7665T>C L1: I426T
g.7677A>C L1: K430T **
Variant type
Gene
M1
M2-M3
M5-M6
M8
Beta
M9
Gamma
Alpha
Variant frequency
1
0.8
0.6
0.4
0.2
0
Gene/region
E1
E2
L1
L2
URR
E1^E4
Intergenic
Variant type
Nonsynonymous
Synonymous
Noncoding
HyPhy MEME P-value
*: P < 0.1
**: P < 0.05
***: P < 0.01
B
HyPhy MEME P-value
*: P <0.1
**: P <0.05
***: P <0.01
①
L1: p.9 A->G ***
L1: p.407 Q->L **
④
E1: p.27 N->K *
E1: p.28 E->K *
E1: p.64 K->T ***
E1: p.139 E->V ***
E1: p.142 L->Q *
E1: p.409 V->C **
L1: p.430 T->K **
②
L1: p.221 P->R *
L1: p.295 P->V *
L1: p.415 E->Q *
Beta-PVs
Gamma-PVs
Alpha-PVs
C
E1
K27N
K28E
T64K
V139E
Q142L
C409V
1
100
200
300
400
500
600
649
Helicase
N-terminal (1-201)
DNA binding domain (203-347)
Minimal oligamerization domain (348-430)
ATPase domain (431-623)
C-terminal (624-649)
D
L1
A9G
P221R
P295V
Q407L
E415Q
K430T
B1
B2
BC-Loop
C
D
DE-Loop
E
DE-Loop
α1
F
FG-Loop
G1
G2
H1
H2
HI-Loop
I
α2
α3
α4
J
α5
1
100
200
300
400
505
N-terminal (1-9)
β-sandwich (27-37, 40-47, 72-80, 96-107, 155-165, 244-255, 300-304, 311-314, 329-338, 342-346, 365-376, 446-455)
α-helix (234-240, 384-393, 396-403, 413-429, 462-473)
Loop (50-69, 110-153, 170-189, 262-291, 348-360)
C-terminal (475-505)
Alpha-specifc
Beta-specific
Beta-and-Gamma-specific


A
Host
Human
Old-world monkeys & apes
New-world monkeys
Non-primate mammals
Reptiles
Birds
Risk
High-risk
Probably high-risk
Low-risk
Unclassified
Non-human
Disease
CSCC
CIN
OPSCC
Benign disease
Immune disorder
Healthy
Non-human
NA
Nucleotide state
A
T
C
G
Gap
Branching point between low-risk and high-risk/probably high-risk HPVs
HPV11
HPV6
HPV18
HPV52
HPV58
HPV35
HPV16
B
Host
Risk
Disease
C
HyPhy aBSREL site-level ω
HyPhy MEME P-value
High-risk
Low-risk
Genome position
Protein position
E1
E2
g.2348C>A p.P245H
g.2412T>G p.S266R
g.2599A>G p.I329A
g.2600T>C p.I329A
g.2653C>G p.Q347E
g.2666A>C p.N351T
g.2674A>C p.T354P P<0.01
g.2680G>A p.E356K
g.3097C>G p.Q495E P<0.05
g.3230C>T p.A539V
g.3278T>C p.V555A
g.3325G>A p.G571S
g.3344C>A p.P577H P<0.1
g.3502A>G p.N630D
g.3518T>G p.L635W
g.3530A>G p.K639R
g.3547A>G p.N645D P<0.01
g.3568T>A p.N21K P<0.05
g.4053T>A p.V183E
g.4089G>C p.C195S
g.323G>A p.S273N
Empirical Bayes factors (EBF)
0.1 1 10 100 1000
Number of substitutions per codon
0 1 2 3
D
HPV16 Ref (with start reset)
E6
E7
E1
E2
E5
L2
L1
Kernel-smoothed frequency
0.075
0.050
0.025
0.000
Genomic position (kb)
0 1 2 3 4 5 6 7 8
Risk-defining signature SNV sites
HPV genome breakpoints (Porter et al., 2024)
E
E1.T354P (CTFELSQMVQWAYDN)
DR
DQ
DP
1/(%rank score)
10
1
0.1
0.01
0.00024
0.058
0.031
DRB1*12:02
SB
WB
High-risk
Low-risk
F
E1.P577H (AGTDSRWPYLHNRLVVFTF)
2.8e−12
5.4e−09
2.1e−05
DRB1*15:02
G
E2.N21K (NVCQDKILTHYENDSTDLR)
0.027
2.8e−08
1.9e−09
DPA1*02:01/DPB1*01:01


A
E1
E2
E6
E7
L2
E1^E4
* Sites under diversifying selection
Cohort 1 (Complete HPV genomes from Genbank, n=194)
Cohort 2 (Partial HPV genomes from Genbank, n=5871)
F1 score
Precision
Recall
cutoff: 0.9
Not covered
1.00
0.75
0.50
0.25
0.00
g.2599A>G
g.2600T>C
g.3502A>G
g.2653C>G
g.2680G>A
g.3278T>C
g.3547A>G
g.3530A>G
g.3518T>G
g.2412T>G
g.1540C>A
g.2666A>C
g.5966A>T
g.4323G>A
g.1115T>G
g.5471C>T
g.1061T>C
g.1144T>A
g.5378C>G
g.3230C>T
g.3344C>A
g.1065C>G
g.3325G>A
g.3097C>G
g.5912A>G
g.4231A>G
g.2674A>C
g.3568T>A
g.5692A>C
g.851G>C
g.4089G>C
g.2348C>A
g.4053T>A
g.6241A>T
B
Cohort 3 (CSCC-HPV16 genomes collected in this study, n=131)
Cohort 4 (OPSCC-HPV16 genomes collected in this study, n=63)
g.851G>C
g.1061T>C
g.1065C>G
g.1115T>G
g.1144T>A
g.2348C>A
g.2412T>G
g.2599A>G
g.2600T>C
g.2653C>G
g.2666A>C
g.2674A>C
g.2680G>A
g.3097C>G
g.3230C>T
g.3278T>C
g.3325G>A
g.3344C>A
g.3502A>G
g.3518T>G
g.3530A>G
g.3568T>A
g.4053T>A
g.4089G>C
g.4231A>G
g.4323G>A
g.5378C>G
g.5471C>T
g.5692A>C
g.5912A>G
g.5966A>T
g.6241A>T
g.1540C>A
g.3547A>G





E

CSCC (n=131)

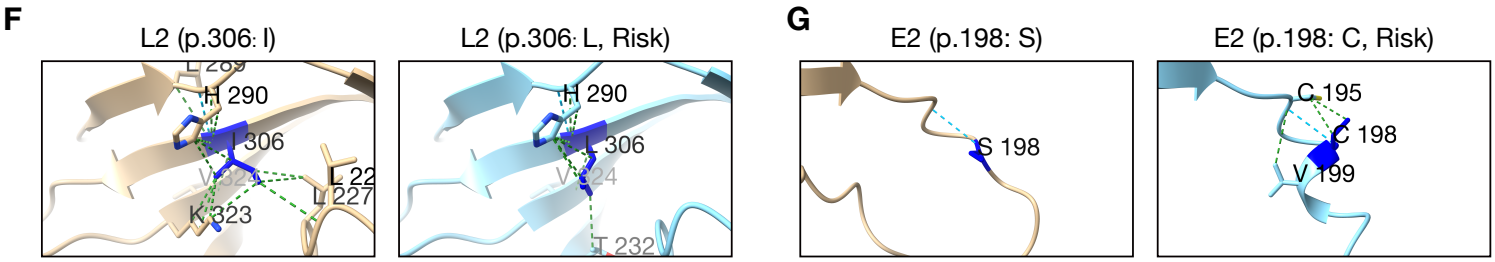
Characteristics	Number (%)	Hazard ratio (95%CI)	*P*-value
Age		1.04 (0.97 - 1.11)	0.25
Stage			
I & II	116 (88.55%)		
III & IV	15 (11.45%)	12.33 (3.21 - 47.36)	<0.001
With prognostic HPV16 variant(s)			
No	126 (96.18%)		
Yes	5 (3.82%)	30.39 (4.99 - 185.17)	<0.001

Global P–value (Log–Rank): <0.001
AIC: 89.77; Concordance Index: 0.78

1 10 150

OPSCC (n=63)

Characteristics	Number (%)	Hazard ratio (95%CI)	*P*-value
Age		0.91 (0.71 - 1.17)	0.48
Stage			
I & II	55 (87.30%)		
III & IV	8 (12.70%)	43.43 (1.49 - 1264.21)	0.028
With prognostic HPV16 variant			
No	58 (92.06%)		
Yes	5 (7.94%)	31.69 (1.26 - 797.63)	0.036

Global P–value (Log–Rank): 0.0046
AIC: 17.3; Concordance Index: 0.97

1 50 1000