Published in final edited form as:
Virus Genes. 2010 June ; 40(3): 329–340. doi:10.1007/s11262-010-0451-1.

Human Alpha and Beta Papillomaviruses Use Different Synonymous Codon Profiles

Nancy M. Cladel[a,b,*], **Alex Bertotto**[a,b], and **Neil D. Christensen**[a,b,c]

[a]Jake Gittlen Cancer Research Foundation, Pennsylvania State University College of Medicine, Hershey, Pennsylvania, 17033, USA

[b]Department of Pathology, Pennsylvania State University College of Medicine, Hershey, Pennsylvania, 17033, USA

[c]Department of Microbiology and Immunology, Pennsylvania State University College of Medicine, Hershey, Pennsylvania, 17033, USA

Abstract

Human papillomaviruses use rare codons relative to their hosts. It has been theorized that this is a mechanism to allow the virus to escape immune surveillance. In the present study we examined the codings of four major genes of 21 human alpha (mucosatropic) viruses and 16 human beta (cutaneous-tropic) viruses. We compared the codon usage of different genes from a given papillomavirus and also the same genes from different papillomaviruses. Our data showed that codon usage was not always uniform between two genes of a given papillomavirus or between the same genes of papillomaviruses from different genera. We speculate as to why this might be and conclude that codon usage in the papillomaviruses may not only play a role in facilitating escape from immune surveillance but may also underlie some of the unanswered questions in the papillomavirus field.

INTRODUCTION

Papillomaviruses are small (about 8Kb) DNA tumor viruses composed of a circular double stranded genome. There more than 100 viruses in this diverse group and each is highly species and tissue specific. A subset of the viruses is associated with cancers, most notably cancer of the cervix, but also head and neck cancers and certain squamous cell carcinomas [1]. Cervical cancer is one of the leading causes of death in women of childbearing age in developing countries and is thus an important public health target [2].

All human papillomaviruses contain the early genes E1, E2, E6 and E7 and the late genes L1 and L2. In addition, there is an E4 gene embedded within the E2 gene; in some viruses, there is also an E5 gene located between the end of E2 and the beginning of L2. E6 and E7 are oncogenes; E1 and E2 are required for replication and gene regulation, and L1 and L2 are capsid genes [3].

Papillomaviruses use rare codons relative to their hosts [4,5]. It has been theorized that this is an evolutionary adaptation that allows the virus to survive within the cell without triggering the immune response [6]. Viral codon usage has also been correlated with tRNA profiles in the cell [7-9]. Adaptation of codon usage to that of the cell infected by the virus

*Corresponding author: Phone: (717)531-6185, Fax: (717) 531-5634, nmc4@psu.edu Nancy M. Cladel .

could help to explain both the tissue and species specificities of these viruses as well as gene expression patterns.

Recent studies have shown that protein expression in *in vitro* systems can be enhanced by codon-modifying the gene to be expressed to match the codon usage of the system in which it is being expressed [10,11]. This technique has greatly increased expression levels *in vitro* and has enabled the production of quantities of proteins that had heretofore been difficult to obtain. We have applied the same technique in our *in vivo* cottontail rabbit papillomavirus (CRPV) animal model to increase protein amounts and to enhance immunogenicity of the viral genes [12]. Our laboratory became interested in codon usage in papillomaviruses based on the findings in these studies.

In the present study the codings of four of the major genes of 21 human alpha viruses and 16 human beta viruses were examined. These viruses represent types within six different species of alpha and two different species of beta viruses as described by deVilliers et al. [13]. Codon usage between genes of a given virus was compared for all amino acids; in selected instances codon usage between the same genes of different genera was examined.

We report here that there were distinct differences in coding patterns between different genes in the same virus and also between the same genes of the alpha and beta papillomaviruses. We postulate that codon usage, in addition to providing a mechanism for escape from immune surveillance, may also yield clues to the questions surrounding issues of tissue and species specificity of the viruses and temporal expression of the viral genes.

MATERIALS AND METHODS

Retrieval of papillomavirus genomes from GenBank

The sequences of the human papillomaviruses used in this study were downloaded from GenBank (www.ncbi.nlm.nih.gov/).

Virus	Accession number
HPV2A	X55964
HPV3	X74462
HPV5	NC_001531.1
HPV6B	NC_001355.1
HPV7	NC_001595.1
HPV8	M12737.1
HPV9	NC_001596
HPV10	NC001576.1
HPV11	EU918768
HPV12	X74466
HPV13	DQ344807.1
HPV14D	X74467.1
HPV15	X74468.1
HPV16	EU918764.1
HPV17	X74469
HPV18	EF202154.1

HPV19	X74470
HPV20	U31778.1
HPV21	U31779.1
HPV22	U31780
HPV23	U31781
HPV24	NC_001683
HPV25	X74471
HPV26	NC_001583
HPV28	U31783.1
HPV29	U31784.1
HPV30	X74474.1
HPV 31	J04353.1
HPV32	NC_001586
HPV33	EU918766.1
HPV 34	NC_001587.1
HPV35	M74117.1
HPV36	U31785.1
HPV37	U31786
HPV38	U31787
HPV39	M62849.1
HPV40	X74478.1
HPV42	M73236.1
HPV43	AJ620205.1
HPV44	U31788.1
HPV 45	EF202166.1
HPV47	M32305.1
HPV48	NC_001690
HPV49	NC_001591.1
HPV50	NC 001691
HPV51	M62877.1
HPV52	X74481
HPV53	NC_001593
HPV54	AF436129
HPV55	U31791
HPV56	EF177181
HPV57	AB361563
HPV58	EU918765.1
HPV59	EU918767.1
HPV60	NC_001693
HPV61	NC_001694.1
HPV65	X70829
HPV66	U31794.1
HPV67	D21208

HPV68	EU918769.1
HPV69	AB027020
HPV70	U21941.
HPV71	AY330621
HPV73	X94165
HPV74	AF436130
HPV75	Y15173.1
HPV76	Y15174.1
HPV77	Y15175.1
HPV80	Y15176.1
HPV 83	AF151983.
HPV92	NC_004500
HPV96	NC_005134

The E1, E2, L1, and L2 genes of the following viruses, chosen at random from the larger set, were loaded into DNAMAN (Lynnon Corporation, Pointe-Claire, Quebec, CANADA).

Alpha viruses: 2a, 3, 6b, 7, 10, 11, 13, 16, 18, 26, 30, 31, 33, 34, 35, 39, 40, 45, 51, 52, 58,

Beta viruses: 5, 8, 9, 12, 14D, 15, 17, 19, 20, 22, 23, 24, 25, 36, 38, 47.

Codon usage for each amino acid for each gene was computed using the DNAMAN program.

Comparison of amino acid codings between genes of a given virus

The codings for a given amino acid and a given gene (E1, E2, L1, or L2) of a given virus were compared with the codings for the same amino acid for each of the other genes for the same virus using Chi Square statistics. Both Smith's Statistical Package *{http://www.economics.pomona.edu/StatSite/SSP.html*) and Calculation for the Chi Square Test [(Preacher, K. J. (2001, April). Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. Available from http://www.quantpsy.org.)} were used for the analyses. The benefit of the latter program is that it gives a Yate's correction for each value. A value of P</=0.05 was considered significant.

Composite codon comparisons for the amino acids serine, threonine, glycine arginine and proline

Both within genus and between genera comparisons of codings for a given gene for amino acids serine, threonine, glycine, arginine and proline were determined. These amino acids were chosen for analysis because they had been shown in the between-gene comparisons above to often differ in their codon usage. To accomplish the within genus analyses, the 21 alpha viruses and 16 beta viruses were divided into arbitrary halves and the codings for each amino acid for each half were summed. Each half was compared with the other half using the Chi Square statistic with Yate's correction.

For the between genus comparisons, the total codings for each amino acid and each gene for the alpha viruses were compared by the Chi Square statistic with the total codings for the same amino acid and gene of the beta viruses.

Calculation of preferred codons for amino acids serine, threonine, glycine, arginine and proline

The composite codings determined above were used to ascertain the preferred codons for the five amino acids serine, threonine, arginine, glycine and proline for the genes E1, E2, L1 and L2 for the alpha and beta papillomaviruses. The number of codons for a given triplet divided by the total number of codons in each composite gene was calculated to yield the percent usage of that codon.

Comparison of lysine coding for L1 for the viruses in the Zhao et al. analysis [14]

The following 72 viruses, representing most of the human viruses investigated in the Zhao study [14] were evaluated for L1 codon usage for lysine.

2A, 3, 5, 6B, 7, 8, 9, 10, 11, 12, 13, 14D, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 65, 66, 67, 68, 69, 70, 71,73, 74, 75, 76, 77, 80, 83, 92, 96.

The codon usage for lysine for the L1 gene for each virus was computed using DNAMAN. The data were compared to that of the Zhao study [14].

Comparison of serine and threonine codings for beta virus E2 genes and for the E2 sequences minus the E4 overlap

Our analysis showed that serine and threonine codings were different in the beta viruses for E2 vs the rest of the genes. We postulated that this could be due to requirements of the overlapping E4 gene. Therefore we used the Chi Square statistic to compare codon usage between the full E2 gene of the beta viruses and the E2 sequence minus the E4 overlap.

Comparison of alpha and beta papillomavirus E2 and E4 genes

Our analysis revealed differences in the lengths of the alpha and beta virus E2 and E4 genes and in the numbers of prolines in the respective E4 genes. We therefore documented these differences in tabular form. Length of each gene, total number of prolines per E4 gene, and numbers of consecutive prolines were compiled.

RESULTS

The early genes, E1 and E2, of the ALPHA papillomaviruses coded differently from the late genes, L1 and L2, for certain amino acids

For the amino acids SERINE, THREONINE, ARGININE, PROLINE, and GLYCINE, the E1 and E2 genes often coded statistically significantly differently from the L1 and L2 genes of the corresponding ALPHA virus. The most significant differences were seen between E1 and L2 for which pair 20 of 21 viruses coded differently for serine, 18 of 21 for arginine, 18 of 21 for proline 15 of 21 for threonine and 14 of 21 for glycine. Most differences were highly significant ($P<0.01$). L1 and L2 for a given virus generally coded the same; there were some differences between E1 and E2; most were not highly significant. The serine analysis was continued for six additional alpha viruses, one from each of a species not represented in the original analysis, to see if the pattern continued to hold. All six viruses showed statistically significant differences between serine coding for E1 vs L2. Thus 26 of 27 alpha papillomaviruses coded statistically differently for E1 and L2 for serine. Table 1 summarizes the data for SERINE. The data for the remaining amino acids may be found as supplemental materials, Tables S1A-D.

The early genes, E1 and E2, of the BETA papillomaviruses coded differently from each other for SERINE and THREONINE. E2 coded differently from L1 and L2 for these amino acids; E1 coded the same

All of the beta papillomaviruses tested (16 of 16) coded differently for SERINE for the protein pair E1/E2. 13 of 16 of the viruses coded differently for serine for E2/L1 and 14 of 16 for E2/L2. Many differences were highly significant ($P<0.001$). E1/L1 and E1/L2 codings were generally the same. L1 and L2 codings were the same for serine for a given virus. When the analysis for serine was continued for three additional beta viruses representative of three additional species, two of the three were found to code significantly differently for E1 and E2. Thus 18 of 19 beta viruses coded differently for serine for E1 and E2. Table 2 summarizes the serine data for the 19 beta viruses.

11 of 16 beta viruses coded differently for THREONINE for the protein pair E1/E2. Many differences were highly significant ($P<0.01$). 10 of 16 of these viruses coded differently for both E2/L1 and E2/L2. There were few differences between E1/L1 and E1/L2. L1 and L2 codings were similar for threonine for a given virus. Thus serine and threonine have similar coding patterns. However, the serine coding differences are greater. The data for threonine may be found in the supplemental materials, Table S2A.

For the remaining amino acids investigated in this analysis, E1 and E2 coded the same with respect to each other in the beta papillomaviruses, as did L1 and L2

E1 and E2 generally coded the same with respect to each other for ARGININE, PROLINE, and GLYCINE, in the beta papillomaviruses studied. The same was true for L1 and L2. Unlike the alpha viruses, in which there were significant differences between E1 and L1/L2, there were fewer differences between E1 and L1 or L2 in the beta viruses studied. The data for the remaining amino acids may be found in the supplemental materials, Tables S2B-D.

The alpha and beta papillomaviruses could be distinguished from each other by their differential codings for serine

Beta viruses coded differently for serine for the protein pair E1/E2 in 18 of 19 viruses. A single beta virus coded significantly differently for the E1/L2 pair and only two for E1/L1 pair. Most beta viruses coded differently for E2 and L1 or L2. On the other hand, 26 of 27 alpha viruses coded differently for the E1/L2 pair and 21/26 for E1/L1 pair. This differential codon usage for serine in proteins of the alpha and beta viruses could be used to discriminate between the two virus genera.

Composite codon analysis confirmed that there were large differences in amino acid codings between the alpha and beta viruses

Our data suggested that codon usage for a given gene differed between the alpha and beta papillomaviruses. The data also suggested that within a genus the codings for a given gene were generally the same. We wanted to verify these findings. In order to do this, we compiled composite codings and compared them to each other by Chi-Square analysis. Tables 3 A, B and C summarize the results of this analysis. Table 3A shows the *within alpha* genus analysis and table 3B, the *within beta* genus analysis. Table 3C shows the *between* genera analysis. Codon usage within groupings of alpha or beta viruses for a given gene was analyzed. For all of the beta papillomaviruses in this study, the within gene differences for these groups were not significant. ($P >0.1$ Chi-Square statistic, with Yate's correction). A few relatively small but significant differences were seen between the alpha papillomavirus groupings ($P>0.01$). Codon usage between groupings of alpha and beta viruses (table 3C) was significantly different for most of the amino acids and genes ($P<0.0004$, Chi-Square analysis after Yate's correction). The exception of note was usage for proline in the E2 gene

where no significant difference was found between the alpha and beta viruses. These data confirmed our earlier data that codon usage for the amino acids investigated in this study was, in general, very different for alpha and beta viruses.

Codon usage for a given amino acid differed from gene to gene within a genus and between genes of the two genera

Table 4 summarizes codon usage for the four genes and five amino acids investigated in this study. This summary was generated from the composite codon analyses. Codons are listed in order of their use in a given gene and it can be seen that this order differs from one gene to another and from one genus to the other.

Codings for beta papillomavirus serines and threonines in the hinge of E2 were correlated with prolines in overlapping E4

Our finding that serine and threonine codings were different for beta virus E2 genes relative to the E1, L1 and L2 genes led us to compare codon usage for the complete E2 gene with usage in the E2 sequence minus the E4 overlap. Composite codon usage analysis using the Chi Square statistic showed that usage for the full E2 gene was statistically the same for the beta viruses; likewise usage for the sequences minus the E4 overlap was the same. On the other hand, composite analysis between the full E2 genes and the sequences minus the E4 overlap was highly significantly different ($P<10^{-8}$). We found that the TCC codings for serine and the ACC codings for threonine were related to the encoding of prolines in E4. These codings were almost exclusively restricted to the area of overlap of the two genes. This demonstrated that while TCC and ACC are not preferred codons in the rest of the genome, they have been selected in the hinge to allow for the encoding of prolines in E4. Table 5 illustrates the composite codon usage for serine and threonine for full length E2 and for and E2 minus E4 for the 16 beta viruses examined in this study.

Alpha and beta papillomaviruses differed in the length of their E2 and E4 genes and in the proline content of the E4 gene

Compilation of all of the alpha and beta virus E2 genes in this study showed that the beta virus genes were longer, on average, than the alpha virus counterparts. The beta virus E4 genes were about twice as long as those of the alpha viruses. They contained, on average, three times as many prolines and these prolines tended to be found more often in clusters of three or more. The data are summarized in Table 6.

Preferred codon for LYSINE for the L1 gene is AAA

As the results of our study began to be assembled, we recognized that our data were not always consistent with those of Zhao et al. [14]. Since our study included only a subset of the viruses in their analysis, we recognized that it would be useful to examine the coding for at least one amino acid in the entire set of viruses examined in that paper. We observed that the codon usage for lysine for L1 in our analysis did not appear to be consistent with their data and chose to focus on this amino acid since it had only two codons and the analysis could be expected to give definitive results. After tabulating codon usage for lysine in L1 for these 72 viruses, we found that AAA was used at a ratio of 2:1 over AAG. These results are the opposite of those found by Zhao et al. [14] who reported that AAG was used twice as often as AAA. Bravo and Muller [15] have also reported disagreements between their codon analysis and that of Zhao et al. [14]. Since the raw material for these data analyses was derived from publicly available databases, the analyses should have yielded the same results. Results of our analysis of lysine codon usage in the L1 gene of the 72 viruses in the Zhao study are shown in Table 7.

DISCUSSION

In this paper, we have shown that codon usage between different genes of the same papillomavirus and between the same genes of alpha and beta viruses is not always the same. Our analysis detected certain amino acids for which differences in codon usage were especially prevalent. The subsequent focus of the study was placed on those amino acids; analysis has been restricted to the early genes E1 and E2 and the late genes L1 and L2 because they are large enough to allow one to adequately assess relative codon usage within single viruses.

Among the major problems still unresolved in papillomavirus research are 1) the mechanism(s) underlying tissue and species specificities, and 2) the means by which early and late genes are differentially regulated. Papillomaviruses are highly tissue and species specific [16]. To date, explanations for these specificities have challenged investigators. One of the ideas that has begun to link the questions of both tissue specificity and gene regulation in living systems is that of the function of synonymous codon usage in contributing to each. Thus, Mukopadhway et al. [17] have shown that different evolutionary forces underlie synonymous codon usage in tissue-specific vs. housekeeping genes of rice and Arabidopsis. Ren et al. [18] have reported that developmental stage-related patterns of gene expression are correlated with CG3 (C or G usage at the third position of the codon) and have found evidence that natural selection acts at synonymous sites in the mouse genome. Dittmar et al. [19] demonstrated the tissue-specific expression of tRNA species, which argues for a role of tRNA heterogeneity in regulation of translation. Zhao et al. [7] found that keratinocytes, the cells in which papillomaviruses carry out their life cycle, experience different tRNA profiles as they differentiate. In a later paper they expanded this work to show that codon usage and tRNA profiles were linked [9]. The work of Kryazhimsky et al. [20] supports selection at synonymous sites for the influenza A virus and suggests that synonymous codon use could be the result of selection for codons that will be translated at the rate optimal for the virus (not necessarily the highest rate). These studies and others [see, for example 21, 22, 23, 24, 25, 26, 27] therefore link either tissue specificity and/or temporal expression with codon usage.

Our studies have shown that for a given **alpha** virus, the late proteins, L1 and L2, normally coded statistically the same. In our examination of codings for all 18 amino acids with two or more possible codons, very few statistical differences were found between L1 and L2 codings (data not shown). With some exceptions, the same pattern was found for the early proteins, E1 and E2, for the alpha viruses. That is, these two proteins generally coded similarly with respect to each other for a given virus. However, when we compared the early protein codings with those of the late proteins, we found that there were many differences. These differences were striking for certain amino acids, the ones that became the focus of this study (Table 1; supplemental data tables S1A-D). The most notable differences were between E1 and L2 codings for which pair E1 coded statistically differently from L2 for serine (26 of 27), arginine (18 of 21), proline (18 of 21) threonine (15 of 21) and glycine (14 of 21). There were fewer differences between E1 and L1 for serine (22 of 27), arginine (13 of /21), proline (11 of 21), threonine (15 of 21) and glycine (10 of 21) indicating that while there were no statistically significant differences in codings between L1 and L2, there were subtle differences that became evident when E1 codings were compared to those of each of the late genes. There were fewer differences between E2 codings and those of the late genes, again demonstrating that differences that were not statistically apparent (as for the E1 and E2 pair or the L1 and L2 pair) sometimes became apparent when comparing the early genes with the late genes.

We suggest that the codon differences may reflect differences in tRNA profiles in cells at different stages of differentiation. E1 and E2 are produced early in the life cycle of the virus. L1 and L2 are produced in terminally differentiating cells. Differential synonymous codon usage could help to determine differences in temporal expression. The patterns that were revealed by this work showed that alpha papillomavirus E1 coding was farther removed from L1 and L2 codings than was E2 coding; furthermore, E1 coding was farther removed from L2 than from L1. These observations could be a clue to the relative temporal expressions of these two early genes as well as to those of the late genes. Ozbun and Meyers [29] reported that E1 and E2 levels were different from each other at different stages of the life cycle of the HPV31b virus with E2 expression being relatively constant and E1 expression varying. Florin et al. [30] reported that L2 is expressed and translocated to the nucleus before L1. We hypothesize that these differential expressions could be regulated by different tRNA pools expressed at different stages of cellular differentiation. The differentials may be subtle, as in the case of the L1/L2 pair, or highly significant as in the case of the E1/L2 pair.

Beta viruses exhibited different codon usage patterns from the alpha viruses. While they generally coded the same for L1 and L2 for a given virus, they coded very differently for E1 and E2 in the case of both serine (18 of 19) and threonine (11 of 16). E1 exhibited the same codings for serine and threonine as did L1 and L2 and thus it was E2 whose codon usage was different. This pattern was the opposite of that seen for the alpha viruses in which 26 of 27 viruses examined coded *differently* for serine for E1/L2. These differences in coding patterns could be used to separate the alpha and beta viruses. They may also provide clues to differences in temporal expression of the beta virus genes *in vivo*. Beta papillomavirus life cycles have not been studied extensively and little information is available on the relative expression of the genes of these cutaneous viruses.

This work demonstrated that the beta viruses, as a group, were different from the alpha viruses, supporting the phylogenetic separation of the two genera [13]. Composite codon analysis showed that for a given amino acid, codon usage for E1, E2, L1 and L2 was highly significantly different between the alpha and beta viruses (Table 3C). Composite codon analysis between viruses *within* the same genus, alpha or beta, did not show the striking differences seen when *between* genera comparisons were made (Table 3A and B). The exceptions in the case of the alpha virus groupings may not be surprising in view of the fact that the alpha papillomaviruses are reported to demonstrate phylogenetic incongruence [31]. These data bolster our findings for the individual virus comparisons and serve to reinforce the conclusion that there are major differences in codings between the alpha and beta papillomaviruses, differences not seen within each genus.

The preferred codon for serine in beta virus E2 genes was found to be TCC and the preferred codon for threonine was ACC. In our analysis of the codings for five amino acids in each of four genes, these were the only preferred codons not ending in A or T (Table 4). The E2 protein is a modular protein consisting of a 5′ transactivation domain (TAD), a 3′ DNA binding domain (DBD) and an internal hinge connecting the two. The TAD and DBD domains are quite highly conserved but the hinges are not. The E4 gene is encoded by the same stretch of DNA in an alternate reading frame and overlaps the hinge [32]. We observed that the beta virus E2 and E4 genes were larger than the alpha virus counterparts. (Table 6). The beta virus E2 genes are serine and threonine-rich and the beta virus E4 genes are proline-rich. It is the presence of the unusual (for papillomaviruses) codons, TCC for serine and ACC for threonine, in E2 that allows for the encoding of prolines in the alternative reading frame E4.

Overlapping genes present an interesting study for evolutionary biologists in that it is challenging to understand how two genes encoded in different frames of the same sequence may be experiencing selection simultaneously. Hughes and Hughes [33] have examined overlapping genes in closely related papillomavirus pairs and have concluded that, with respect to E2 and E4, the E2 hinge is evolving under diversifying selection and E4 is evolving under purifying selection. Similar results were reported in the same year by Narechania et al. [34. In light of these findings, we suggest that proline in beta papillomaviruses is under purifying selection in the E4 gene and that the unusual codings for serine and threonine in E2 (relative to the rest of the virus) have been selected to allow for these prolines in E4. In support of this theory, our analysis showed that the codings for serine (TCC) and threonine (ACC) in E2 were predominantly confined to the hinge region overlapping E4.

Whereas the E4 genes of alpha viruses have been studied extensively [31], those of the beta papillomaviruses have received little attention. Our investigations have shown that these genes are considerably longer than those of the alpha viruses and that they contain multiple polyproline regions (Table 6). Williamson [35], in a review on proline-rich regions in proteins, noted that proteins with abundant amounts of this amino acid bind non-stoichiometrically but functionally to other proteins. It remains to be determined if this is true for the papillomavirus E4 proteins and, if so, what the function might be. This area may be an important avenue for future investigations. Prolines in the alpha virus E4 genes tend to occur singly or in pairs, whereas, the beta virus E4 genes all contain at least one run of three consecutive prolines and often contain runs of five or more. We postulate that these differences could be related to the tissue-specific differences of the two genera. Recent work on keratinocyte proline-rich protein (KPRP) in several laboratories [36, 37] discusses the association of this protein with the cornified envelope in terminally differentiating keratinocytes. We wonder if the beta virus E4 proteins, which are especially proline rich, might subvert this association with an interaction of their own to foster the ultimate release of the virus.

The only other in depth study of codon usage in papillomaviruses was done by Zhao et al. [14]. In that study, alpha, beta, gamma and mu human papillomaviruses were combined and more than 70 viruses were analyzed for codon usage. Because the different genera were combined, the differences we have noted between the alpha and beta viruses were not detected. In addition, the conclusion that T was favored over A at the third position of papillomavirus codons did not seem justified to us. We checked the coding for lysine in L1 for all of the viruses in their study and found that AAA was favored by a factor of 2:1 over AAG. This was the opposite of what was reported by these investigators. In support of our findings, Bravo and Muller [15] have also reported that the data of Zhao et al [14] are not consistent with published databases.

In summary, this study of existing data extracted from GenBank and mined for codon usage has shown that genes of a given papillomavirus do not always code the same. The differential codings may provide clues to the mechanisms underlying temporal differences in expression of viral genes and tissue specific expression of the papillomaviruses. While differences between our data and that of Zhao et al [14] were found, our work does lend support to the findings of that laboratory that codon usage and tRNA profiles are linked [7,8].

We have demonstrated differences in codings for serine and threonine in the beta virus E2 genes relative to the other viral genes. These differences were linked to the encoding of prolines in the overlapping E4 gene. In addition, we have shown that alpha and beta viruses use different codons for serine for E2. Alpha viruses are primarily mucosal and beta viruses

are universally cutaneous in their tissue tropism. Their respective E4 genes are very different, both in length and in proline content. Our findings could pave the way for the study of beta virus E4 genes and may also provide a place to focus exploration of the differences in tissue specificities of the alpha and beta viruses.

These observations could also further studies on selection in overlapping genes, a topic of considerable current interest. The beta papillomaviruses, in particular, may prove to be useful tools to investigate this question since our data supports selection of synonymous codons of serine(TCC) and threonine(ACC) in the E2 gene to allow for the encoding of prolines in the overlapping E4 gene.

Finally, most phylogenetic analyses eliminate the third position of a codon due to the assumption that this data is saturated [38, 39]. This simplifies the analyses and has generally given rise to well-substantiated trees. On the basis of the data generated here, we suggest that the inclusion of the third position in follow-up phylogenetic analyses could help to further refine the topology, especially in the cases of closely related viruses. This thought is supported by the work of Ren et al. [40], Yang et al. [39] and Bodilis and Barray [38].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Cancer Institute grant R01 CA47622 from the National Institutes of Health and the Jake Gittlen Memorial Golf Tournament.

Reference List

1. Kreimer AR, Clifford GM, Boyle P, Franceschi S. Cancer Epidemiology Biomarkers & Prevention. 2005; 14:467–475.
2. zur Hausen H. J. Natl. Cancer Inst. 2000; 92:690–698. [PubMed: 10793105]
3. Zheng ZM, Baker CC. Front Biosci. 2006; 11:2286–2302. [PubMed: 16720315]
4. Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. Journal of Virology. 1999; 73:4972–4982. [PubMed: 10233959]
5. Shackelton LA, Parrish CR, Holmes EC. J. Mol. Evol. 2006; 62:551–563. [PubMed: 16557338]
6. Tindle RW. Nat. Rev. Cancer. 2002; 2:59–65. [PubMed: 11902586]
7. Zhao KN, Gu W, Fang NX, Saunders NA, Frazer IH. Mol. Cell Biol. 2005; 25:8643–8655. [PubMed: 16166644]
8. Gu W, Li M, Zhao WM, Fang NX, Bu S, Frazer IH, Zhao KN. Nucleic Acids Res. 2004; 32:4448–4461. [PubMed: 15319446]
9. Gu W, Ding J, Wang X, de Kluyver RL, Saunders NA, Frazer IH, Zhao KN. Nucleic Acids Res. 2007; 35:4820–4832. [PubMed: 17621583]
10. Gao F, Li Y, Decker JM, Peyerl FW, Bibollet-Ruche F, Rodenburg CM, Chen Y, Shaw DR, Allen S, Musonda R, Shaw GM, Zajac AJ, Letvin N, Hahn BH. Hum. Retroviruses. 2003; 19:817–823.
11. Mossadegh N, Gissmann L, Muller M, Zentgraf H, Alonso A, Tomakidi P. Virology. 2004; 326:57–66. [PubMed: 15262495]
12. Cladel NM, Hu J, Balogh KK, Christensen ND. PLoS ONE. 2008; 3:e2947. [PubMed: 18698362]
13. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. Virology. 2004; 324:17–27. [PubMed: 15183049]
14. Zhao KN, Liu WJ, Frazer IH. Virus Res. 2003; 98:95–104. [PubMed: 14659556]
15. Bravo IG, Muller M. Papillomavirus Report. 2005; 16:63–72.

16. Stanley MA, Pett MR, Coleman N. Biochem. Soc. Trans. 2007; 35:1456–1460. [PubMed: 18031245]
17. Mukopadhway P, Basak S, Ghosh TC. DNA Research, advance access. 2008; 15(6):347–356.
18. Ren LC, Gao G, Zhao DX, Ding MX, Luo JC, Deng HK. Genome Biology. 2007; 8
19. Dittmar KA, Goodenbour JM, Pan T. PLoS Genet. 2006; 2:e221. [PubMed: 17194224]
20. Kryazhimskiy S, Bazykin GA, Dushoff J. Journal of Virology. 2008; 82:4938–4945. [PubMed: 18321967]
21. Wang X, Li B, Zhao KN. Virol. J. 2007; 4:127. [PubMed: 18036255]
22. Horn D. BMC Genomics. 2008; 9:2. [PubMed: 18173843]
23. Christianson ML. American Journal of Botany. 2005; 92:1221–1233. [PubMed: 21646144]
24. Akashi H. Current Opinion in Genetics and Development. 2001; 11:660–666. [PubMed: 11682310]
25. Lucks J, Kudla G, Nelson D, Plotkin JB. PLoS Comput. Biol. 2008; 4:e1000001. [PubMed: 18463708]
26. van Hemert FJ, Berkhout B, Lukashov VV. Virology. 2007; 361:447–454. [PubMed: 17188318]
27. Kanduc D. Archives of Biochemistry and Biophysics. 1997; 342:1–5. [PubMed: 9185607]
28. Elf J, Nilsson D, Tenson T, Ehrenberg M. Science. 2003; 300:1718–1722. 13. [PubMed: 12805541]
29. Ozbun MA, Meyers C. Virology. 1998; 248:218–230. [PubMed: 9721231]
30. Florin L, Sapp C, Streeck RE, Sapp M. J. Virol. 2002; 76:10009–10014. [PubMed: 12208977]
31. Narechania A, Chen Z, DeSalle R, Burk RD. J. Virol. 2005; 79:15503–15510. [PubMed: 16306621]
32. Roberts, S. Biology of the E4 protein. In: McCance, DJ., editor. Human Papillomaviruses. Vol. 8. Elsevier Science B.V.; Amsterdam, The Netherlands: 2002. p. 119-142.
33. Hughes AL, Hughes MA. Virus Res. 2005; 113:81–88. [PubMed: 15913825]
34. Narechania A, Terai M, Burk RD. J. Gen. Virol. 2005; 86:1307–1313. [PubMed: 15831941]
35. Williamson MP. J. 1994; 297:249–260.
36. Kong W, Longaker MT, Lorentz HP. J. Biol. Chem. 2003; 278:22781–22786. [PubMed: 12668678]
37. Lee W, et al. J of Invest Dermatol. 2005; 125:995–1000. [PubMed: 16297201]
38. Bodilis J, Barray S. Microbiology. 2006; 152:1075–1088. [PubMed: 16549671]
39. Yang X, Tuskan GA, Tschaplinski TJ, Cheng Max Z. Nature Preceedings. 2007; 10:1038.
40. Ren FR, Tanaka H, Yang ZH. Systematic Biology. 2008; 54:808–818. [PubMed: 16243764]

Table 1

Comparison of serine codon usage between alpha papillomavirus gene pairs. Significance was determined by Chi Square analysis as per Materials and Methods. P</=0.05 was considered significant but all P values are indicated.

PV	E1/E2	L1/L2	E1/L1	E1/L2	E2/L1	E2/L2
16	0.14	0.79	0.01	0.02	0.48	0.58
31	0.48	0.85	0.21	0.03	0.85	0.37
33	0.3	0.34	0.01	0.002	0.54	0.49
35	0.37	0.43	0.001	0.004	0.15	0.58
52	0.03	0.62	0.02	0.003	0.19	0.1
58	0.08	0.23	0.007	0.000001	0.3	0.003
6	0.18	0.74	0.18	0.09	0.28	0.15
11	0.29	0.25	0.06	0.004	0.15	0.52
13	0.07	0.23	0.02	0.001	0.03	0.33
34	0.19	0.75	0.07	0.005	0.03	0.003
18	0.41	0.29	0.0003	0.000009	0.01	0.005
39	0.08	0.3	0.001	0.003	0.01	0.2
45	0.48	0.44	0.02	0.001	0.02	0.01
26	0.89	0.14	0.00003	0.0003	0.02	0.04
30	0.002	0.22	0.02	0.002	0.12	0.8
51	0.51	0.81	0.002	0.00003	0.11	0.01
2a	0.09	0.46	0.0003	0.03	0.07	0.54
3	0.02	0.89	0.0001	0.003	0.37	0.69
7	0.42	0.7	0.003	0.008	0.35	0.09
10	0.01	0.44	0.00007	0.004	0.19	0.7
40	0.43	0.53	0.02	0.01	0.08	0.1
32	0.01	0.32	0.004	0.002	0.36	0.99
83	0.03	1	0.00004	0.00001	0.01	0.009
73	0.17	0.003	0.0002	0.0002	0.26	0.00009
26	0.89	0.14	0.00002	0.0003	0.02	0.04
71	0.04	0.87	0.05	0.01	0.56	0.69

PV	E1/E2	L1/L2	E1/L1	E1/L2	E2/L1	E2/L2
54	0.84	0.26	0.86	0.03	1	0.38

Table 2

Comparison of serine codon usage between beta papillomavirus gene pairs. Significance was determined by Chi Square analysis as per Materials and Methods P</=0.05 was considered significant but all P values are indicated.

PV	E1/E2	L1/L2	E1/L1	E1/L2	E2/L1	E2/L2
5	0.00003	0.61	0.14	0.39	0.0005	0.0003
8	0.003	0.6	0.8	0.62	0.33	0.003
9	0.0001	0.12	0.11	0.46	0.005	1E-08
12	0.0004	0.55	0.47	0.38	0.0009	0.02
14d	1E-08	0.48	0.009	0.26	0.004	1E-08
15	0.0002	0.23	0.42	0.86	0.0005	0.0001
17	0.001	0.47	0.38	0.57	0.005	0.008
19	0.0004	0.63	0.48	0.38	0.04	0.002
20	0.001	0.41	0.76	0.69	0.02	0.002
22	0.03	0.21	0.33	0.48	0.03	0.2
23	0.0007	0.33	0.1	0.09	0.08	0.28
24	0.02	0.44	0.92	0.87	0.1	0.004
25	0.007	0.55	0.77	0.67	0.01	0.0009
36	0.001	0.24	0.93	0.22	0.006	0.02
38	0.0002	0.26	0.23	0.1	0.0003	0.01
47	0.009	1	0.74	0.03	0.02	0.0003
75	0.00004	0.02	0.1	0.7	0.0005	0.001
96	0.09	0.09	0.01	0.55	0.1	0.05
92	0.009	0.56	0.26	0.94	0.004	0.03

Tables 3 A, B, C

Comparison of composite codon usage for genes E1, E2, L1, L2 in A) alpha viruses, B) beta viruses and C) alpha vs. beta viruses for the amino acids noted. Few significant differences were observed in the within genus comparisons (A and B); in contrast, the between genus comparisons(C) showed many significant differences.

Alpha vs. Alpha				
•	**El•**	**E2•**	**LI•**	**L2•**
Arginine•	Chi.•Sq= 13.6• •P = 0.02•	Chi.•Sq=.65• P = 0.99•	Chi.•Sq= 5.8• P = 0.36•	Chi.•Sq= 6.5• P = 0.26•
•	•	•	•	•
Glycine•	Chi.•Sq= 1.4• P = 0.84•	Chi.•Sq= 6.4• P = 0.09•	Chi.•Sq= 1.8• P = 0.63•	Chi.•Sq=.35• P = 0.95•
•	•	•	•	•
Proline•	Chi.•Sq=8.7• P = 0.03•	Chi.•Sq=8.6• P = 0.04•	Chi.•Sq= 1.7• P = 0.63•	Chi.•Sq= 11.1• P = 0.01•
•	•	•	•	•
Serine•	Chi.•Sq= 11.7• P = 0.03•	Chi.•Sq=5.1• P = 0.41•	Chi.•Sq= 6.8• P = 0.24•	Chi.•Sq=6.3• p=0.28•
•	•	•	•	•
Threonine•	Chi.•Sq=4.8• P = 0.19•	Chi.•Sq= 7.7• P = 0.05•	Chi.•Sq= 2.0• P = 0.58•	Chi.•Sq=3.1• P = 0.38•
Beta vs. Beta				
•	**El•**	**E2•**	**LI•**	**L2•**
Arginine•	Chi.•Sq= 2.8• •P = 0.72•	Chi.•Sq=4.4• P = 0.5•	Chi.•Sq=8.3• P = 0.14•	Chi.•Sq= 7.6• P = 0.18•
•	•	•	•	•
Glycine•	Chi.•Sq= 3.6• P = 0.31 •	Chi.•Sq=.74• P = 0.86•	Chi.•Sq= 2.8• P = 0.43•	Chi.•Sq=.33• P = 0.95•
•	•	•	•	•
Proline•	Chi.•Sq=.28• P = 0.96•	Chi.•Sq=.63• P = 0.89•	Chi.•Sq=4.3• P = 0.23•	Chi.•Sq= 3.0• P = 0.38•
•	•	•	•	•
Serine•	Chi.•Sq=2.3• P = 0.81•	Chi.•Sq= 2.6• P = 0.77•	Chi.•Sq= 6.8• P = 0.24•	Chi.Sq= 6.0• p=0.31•
•	•	•	•	•
Threonine•	Chi.•Sq= 3.4• P = 0.33•	Chi.•Sq=4.8• P = 0.19•	Chi.•Sq= 3.0• P = 0.39•	Chi.•Sq=4.0• P = 0.26•
Alpha vs. Beta				
•	**El•**	**E2•**	**LI•**	**L2•**
Arginine•	Chi.•Sq= 11.2• •P = .01•	Chi.•Sq= 40.8• P = l×l0−7•	Chi.•Sq= 125.4• P = 0•	Chi.•Sq= 103.4• P = 0•
•	•	•	•	•
Glycine•	Chi.•Sq=34.2• P =l.8×10−7•	Chi.•Sq= 38.7• P = 2×10−8•	Chi.•Sq=34.3• P =l.7×10−7•	Chi.•Sq=8.9• P = .03•
•	•	•	•	•
Proline•	Chi.•Sq=46.1• P = 0•	Chi.•Sq= 2.5• P = .48•	Chi.•Sq=50• P = 0•	Chi.•Sq= 17.1• P =6.6×10−4•

Alpha vs. Alpha				
•	**El•**	**E2•**	**LI•**	**L2•**
•	•	•	•	•
Serine•	Chi.•Sq= 130.8• P = 0•	Chi.•Sq= 144.2• P = 0•	Chi.•Sq=37.6• P = 5×l0−7•	Chi.Sq= 50.5• p=0•
•	•	•	•	•
Threonine•	Chi.•Sq=70.3• P = 0•	Chi.•Sq=90.4• P = 0•	Chi.•Sq= 17.2• P = 6×l0−4•	Chi.•Sq= 13.6• P = 3.4×10−3•

Table 4

Codon usage in order of preference for serine, threonine, arginine, proline and glycine in alpha and beta papillomavirus genes E1, E2, L1 and L2.

SERINE	**E1ALPHA**	**E1BETA**	ARGININE	**E1ALPHA**	**E1BETA**
	AGT(41.7%)	TCT(27.4%)		AGA(47.3%)	AGA(45.9%)
	TCA(21.1%)	AGT(24.6%)		AGG(20.7%)	CGA(17.9%)
	AGC(20.4%)	TCA(23.2%)		CGA(14.4%)	AGG(16.3%)
	TCT (8.3%)	AGC(15.4%)		CGG(7.3%)	CGT(9.6%)
	TCC(4.8%)	TCC(5.5%)		CGT(6.8%)	CGC(5.7%
	TCG(C.6%)	TCG(4.2%)		CGC(3.6%)	CGG(4.5%)
SERINE	**E2ALPHA**	**E2BETA**	ARGININE	**E2ALPHA**	**E2BETA**
	AGT(35.5%)0	TCC(30.3%)		AGA(32.8%)	AGA(29%)
	TCA(21.1%)	TCA(28.4%)		CGA(22.8%)	AGG(24.3%)
	TCC(17.8%)	AGC(18.1%)		AGG(16.2%)	CGA(19.5%)
	TCT(16%	TCT(11.6%)		CGT(12.2%)	CGG(16%)
	AGC(4.6%)	AGT(6.3%)		CGG(11.2%)	CGT(6.1%)
	TCG(4%)	TCG(5.4%)		CGC(6.7%)	CGC(5.2%)
SERINE	**L1ALPHA**	**L1BETA**	ARGININE	**L1ALPHA**	**L1BETA**
	TCT(40%)	TCT (32.6%)		CGT(28.6%)	AGA(45.2%)
	AGT(27.7%)	TCA(24.5%)		AGG(27%)	AGG(21.9%)
	TCC(14.5%)	AGT(23.7%)		AGA(20.1%)	CGC(10.5%)
	TCA(12.5%)	TCC(8.6%)		CGG(10.9%)	CGT(9.7%)
	AGC(4.5%)	AGC(7.6%)		CGC(6.9%)	CGA(8.9%)
	TCG(2.8%)	TCG(2.9%)		CGA(6.5%)	CGG(3.7%)
SERINE	**L2ALPHA**	**L2BETA**	ARGININE	**L2ALPHA**	**L2BETA**
	TCT(45.8%)	TCT(28.8%)		CGT(36.6%)	AGA(34%)
	AGT(25.8%)	AGT(26.0%)		CGC(21.3%)	AGG(18.7%)
	TCC(16%)	TCA(19.1%)		AGG(17.3%)	CGT(17.4%)
	TCA(16%)	AGC(14.0%)		AGA(11.6%)	CGC(15%)
	AGC(5.8%)	TCC(9.1%)		CGG(7.9%)	CGA(9.2%)
	TCG(4.9%)	TCG(3.4%)		CGA(5.3%)	CGG(5.7%)
THREONINE	**E1ALPHA**	**E1BETA**	PROLINE	**E1ALPHA**	**E1BETA**
	ACA(62.1%)	ACA(52.4%)		CCA(67.7%)	CCT(41.7%)

	ACT(16.2%)	ACT(31.2%)			CCT(14.1%)	CCA(40.4%)
	ACG(13.9%)	ACC(11.2%)			CCG(10.6%)	CCG(9.4%)
	ACC(7.7%)	ACG(5.2%)			CCC(7.6%)	CCC(8.8%)
	THREONINE	**E2ALPHA**	**E2BETA**	PROLINE	**E2ALPHA**	**E2BETA**
		ACA(46%)	ACC(43.6%)		CCA(40.8%)	CCA)32.7%)
		ACT(24.4%)	ACA(27.3%)		CCC(26.1%)	CCT(30.8%)
		ACC(22.5%)	ACT(21.7%)		CCT(21.4%)	CCC(27.5%)
		ACG(7.2%)	ACG(7.4%)		CCG(11.8%)	CCG(9.1%)
	THREONINE	**L1ALPHA**	**L1BETA**	PROLINE	**L1ALPHA**	**L1BETA**
		ACA(48.8%)	ACA(46.7%)		CCT(49.1%)	CCA(44.6%)
		ACT(29.2%)	ACT(36.3%)		CCA(27.5%)	CCT(43.8%)
		ACC(18.2%)	ACG(8.6%)		CCC(17.7%)	CCC(7.9%
		ACG(3.8%)	ACC(8.4%)		CCG(5.7%)	CCG(3.7%)
	THREONINE	**L2ALPHA**	**L2BETA**	PROLINE	**L2ALPHA**	**L2BETA**
		ACA(45.2%)	ACA(48.5%)		CCT(55.5%)	CCT(51.3%)
		ACT(32%)	ACT(35.1%)		CCA(23.7%)	CCA(32.5%)
		ACC(17.2%)	ACC(13/6%)		CCC(17%)	CCC(13.5%)
		ACG(5.6%)	ACG(2/8%)		CCG(3.7%)	CCG(2.7%)
	GLYCINE	**E1ALPHA**	**E1BETA**	GLYCINE	**L1ALPHA**	**L1BETA**
		GGA(34.9%)	GGA(41.5%)		GGA(30.2%)	GGT(37.55)
		GGG(25.2%)	GGT(26%)		GGC(25.8%)	GGA(28.2%)
		GGT(21.1%)	GGG(20.5%)		GGT(23%)	GGC(22.6%)
		GGC(18.9%)	GGC(11.9%)		GGG(21%)	GGG(11. 7%)
	GLYCINE	**E2ALPHA**	**E2BETA**	GLYCINE	**L1BETA**	**L2BETA**
		GGA(36%)	GGA(37.8%)		GGT(42.8%)	GGT(40.5%)
		GGT(28.6%)	GGG(29.7%)		GGC(22.3%)	GGC(22.3%)
		GGC(18.4%)	GGC(17.3%)		GGG(21.1%)	GGA(20.8%)
		GGG(17.1%)	GGT(15.6%)		GGA(15%)	GGG(16.4%)

Table 5

Composite serine and threonine codon usage in beta papillomavirus full E2 and E2 minus E4 sequences. Analysis was performed for the 16 beta viruses in this study. TCC for serine was used four times as often in full E2 relative to the E2 minus the E4 sequences; ACC for threonine was used three times as often in the full E2 sequence relative to E2 minus E4 sequences. Codon usage differences are highly statistically significant for serine (p=0, chi-square =114.3 for 5 degrees of freedom) and threonine (p=0, chi-square =53.9 for three degrees of freedom).

CODON	**SERINE E2 %**		**CODON**	**E2minusE4 %**	
TCT	138	14	TCT	36	14
TCC	313	32	TCC	21	8
TCA	218	22	TCA	50	20
TCG	82	8	TCG	13	5
AGT	116	12	AGT	67	26
AGC	119	12	AGC	67	26
	986			254	
CODON	THR E2 %		CODON	E2minusE4 %	
ACT	162	25	ACT	81	36
ACC	266	41	ACC	31	14
ACA	179	27	ACA	91	41
ACG	47	7	ACG	21	9
total	654		total	224	

Virus Genes. Author manuscript; available in PMC 2013 August 25.

Table 6

Tabulation of the length in base pairs(bp) of E2 and E4 genes in the 21 alpha and 16 beta viruses in this study. P = “Proline” where P (1) indicates the number of single prolines in the E4 gene, P (2) indicates the number of times proline occurs in runs of 2, P (3) indicates the number of times the amino acid occurs in runs of 3, P (4) indicates the number of times it occurs in runs of 4 and P (5 or >) refers to the number of times proline occurs in runs of five or more. Total P is the total number of prolines in a given E4 gene.

ALPHA VIRUSES								
PV	**E2bp**	**E4bp**	**P(1)**	**P(2)**	**P(3)**	**P(4)**	**P(5or>)**	**total P**
2A	1176	37 5	14	2	0	0	0	18
3	1152	399	8	2	1	0	0	15
6B	1108	324	8	1	1	0	0	13
7	1128	306	6	5	0	0	0	16
10	1131	330	7	3	0	0	0	13
11	1104	330	7	2	1	0	0	14
13	1134	327	6	0	1	0	1	14
16	1098	291	11	1	0	0	0	13
18	1098	285	7	1	0	0	0	9
26	1128	321	12	3	0	0	0	18
30	1137	321	14	2	0	0	1	23
31	1111	244	9	1	0	0	0	11
33	1062	288	7	1	0	0	0	9
34	1038	327	5	0	0	0	0	5
35	1104	294	11	1	0	0	0	13
39	1113	256	11	2	0	0	0	15
40	1113	360	8	3	1	0	0	17
45	1107	267	7	2	0	0	0	11
51	1077	275	10	2	0	0	0	14
52	1107	309	8	2	0	0	0	12
58	1077	252	5	2	0	0	0	9
	ave.= 1109	ave. = 309						ave=13
BETA VIRUSES								
PV	E2bp	E4bp	P(1)	P(2)	P(3)	P(4)	P(5or>)	total P
5	1545	737	21	2	2	0	3	49

ALPHA VIRUSES								
PV	**E2bp**	**E4bp**	**P(1)**	**P(2)**	**P(3)**	**P(4)**	**P(5or>)**	**total P**
8	1497	690	21	1	3	1	2	48
9	1386	642	23	2	4	0	0	39
12	1485	677	18	3	2	0	2	42
14D	1452	642	18	1	2	0	3	44
17	1359	618	17	4	1	1	0	32
15	1149	696	15	4	2	1	0	34
19	1482	687	17	5	3	1	1	45
20	1494	687	15	0	2	2	3	44
22	1311	582	17	3	1	0	0	26
23	1296	728	15	5	2	0	0	31
24	1404	669	23	4	1	0	1	42
25	1509	702	23	2	3	3	0	48
36	1530	726	25	3	2	0	3	58
38	1326	570	13	4	3	0	0	27
47	1521	668	12	4	2	0	2	39
	ave. = 1422	ave. = 670						ave=40

Table 7

Lysine codings for the L1 gene for the 72 viruses in the analysis of Zhao et. al. [14]. AAA is used 62% of the time and AAG 38% of the time. Papillomavirus genera are represented by A=alpha, B=beta, G=gamma and M=mu.

virus	%K(AAA)	%K(AAG)	Genus	virus	%K(AAA)	%A(AAG)	Genus
2A	27	73	A	40	52	48	A
3	55	45	A	42	58	42	A
5	63	37	A	43	68	32	A
6B	57	43	A	44	52	48	A
7	73	27	A	45	44	56	A
8	55	45	B	47	67	33	B
9	40	60	B	48	62	38	G
10	44	56	A	49	59	41	B
11	58	42	A	50	67	33	G
12	78	22	B	51	59	41	A
13	65	35	A	52	56	44	A
14D	63	37	B	53	61	39	A
15	72	28	B	54	55	45	A
16	79	21	A	55	59	41	A
17	79	21	B	56	65	35	A
18	48	52	A	57	42	58	A
19	70	30	B	58	69	31	A
20	78	22	B	59	59	41	A
21	75	25	B	60	90	10	G
22	84	16	B	61	41	69	A
23	64	36	B	65	81	19	G
24	77	23	B	66	58	42	A
25	85	15	B	67	65	35	A
26	74	26	A	68	59	41	A
28	58	42	A	69	79	21	A
29	42	58	A	70	37	63	A
30	63	37	A	71	65	35	A
31	85	15	A	73	63	38	A

virus	%K(AAA)	%K(AAG)	Genus	virus	%K(AAA)	%A(AAG)		Genus
32	54	46	A	74	52	48		A
33	78	22	A	75	75	25		A
34	56	44	A	76	61	39		A
35	76	24	A	77	42	58		A
36	56	44	B	80	85	15		B
37	74	26	B	83	35	65		A
38	88	13	B	92	54	46		B
39	39	61	A	96	61	39		B