Biomedical Genetics and Genomics

Research Article ISSN: 2398-5399

Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect

Justin B. Miller, Ariel A. Hippen, Sage M. Wright, Caroline Morris and Perry G. Ridge*

Department of Biology, Brigham Young University, Provo, Utah 84602, USA

Abstract

It is well-documented that codon usage biases affect gene translational efficiency; however, it is less known if viruses share their host's codon usage motifs. We determined that human-infecting viruses share similar codon usage biases as proteins that are expressed in tissues the viruses infect. By performing 7,052,621 pairwise comparisons of genes from humans versus genes from 113 viruses that infect humans, we determined which codon usage motifs were most highly correlated. We found that 16 viruses averaged a significant correlation in codon usage with over 500 human genes per viral gene, 58 viruses were highly correlated with an average of at least 100 human genes per viral gene, and 37 viruses were significantly correlated with an average of at least one human gene per viral gene at an alpha level of 7.09 x (0.05 alpha / 7,052,621 comparisons). Only two viruses were not highly correlated with an average of one human gene per viral gene. While relatively few of the interactions were previously documented, the high statistical correlations suggest that researchers may be able to determine which tissues a virus is most likely to infect by analyzing codon usage biases.

Introduction

Amino acids are encoded by DNA triplets known as codons; however, since there are only 20 canonical amino acids and 64 possible codons, multiple codons encode a single amino acid [1]. The majority of amino acids are encoded by 2-6 different codons. Despite multiple codons encoding a single amino acid, codon usage is not random in most species [2-5]. Various species, including many plant species, *E. coli* and *Drosophila*, also maintain DNA triplet preferences, or codon usage biases, over time in both intronic and exonic regions [6-8].

It is generally accepted that non-random mutations occur more frequently at the third position in the codon, and codon bias persists through selection [9,10]. Numerous biological factors create evolutionary pressure to use certain codons. First, an incomplete set of transfer RNAs (tRNAs) or unequal expression of tRNA anticodons within a tissue or species creates pressure for codons with complementary tRNAs available. Second, translational speed may either increase or decrease depending on the codon used, creating pressure to select codons for which translational efficiency matches the needs of the tissue/cell (i.e. suboptimal codons might be preferential to some species for increased translational efficiency, while in other instances suboptimal codons might decrease translational efficiency) [10,11]. Finally, codon usage bias primarily affects the translation of a gene and is a main determinant of gene expression [12].

Recently, significant correlations for codon usage preferences between RNA viruses (e.g. SBV and KV) and their host, the honeybee, were reported [13]. They proposed that such similarities resulted from co-evolution, which typically occurs in a leapfrog fashion (i.e. as the host evolves to combat the parasite, the parasite evolves to adapt to the new conditions).

We aimed to determine whether the same relationship exists between human and viral genes expressed in tissues targeted by the virus. We analyzed 19,482 human proteins, and compared their codon usage biases against 113 viruses that infect human hosts. We found significant correlations for many viral and human proteins, and where tissue information was available, the top correlated human protein was frequently highly expressed in the tissue type targeted by the virus.

Materials and methods

Data collection and cleaning

We used gene annotations from the General Feature Format (GFF) and GFF3 files from the National Center for Biotechnology Information (NCBI) to extract the reference viral and human sequences [14-16]. Since the reference genome is intended to most accurately represent an average individual in a species, we downloaded all reference sequence data, including the corresponding gene annotations, from NCBI. Similar to the methods used by [17], when multiple isoforms were annotated, the longest isoform was always chosen as the representative isoform for that gene, and we removed all genes with any annotated translational exceptions (e.g., translational, unclassified transcription discrepancy, suspected errors, etc.). These filters had only a minor effect on our data because they eliminated less than 5% of the total sequences. All 19,482 sequence accession numbers can be found in the NCBI database by downloading the complete genome annotations for Homo sapiens; the accession numbers for each virus and their highest correlating genes are located in Table S1.

Codon usage correlation values

To determine if there was a correlation between human and viral codon usage biases, we performed a Pearson's r correlation test with

Correspondence to: Perry G. Ridge, Department of Biology, Brigham Young University, Provo, Utah 84602, USA, E-mail: perry.ridge@byu.edu

Key words: *codon usage bias, host, human, virus, virus-host interactions*

Received: June 10, 2017; **Accepted:** July 24, 2017; **Published:** July 27, 2017

discrete codon usage counts by comparing total codon usage counts in human and viral coding sequences (CDS). We used Pearson's r because it uses a product-moment correlation coefficient that is used to determine the correlation between two variables with different units or different magnitudes [18]. Since gene lengths can vary greatly between genes, and genes do not contain all codons, the assumptions for most statistical tools would not be adequately met using the raw data. Furthermore, the high number of zero codon usage counts in some genes meant that a percentage comparison of codon usages using a traditional t-test was unfeasible, even with a transformation. We chose an implementation of Pearson's r from the package SciPy in Python version 2.7 because Pearson's r is robust to variations in sequence sizes as well as zero values. Using Pearson's r, we graphed a linear regression and calculated the R^2 coefficients of determination and p-values by plotting the discrete codon counts from each gene within each virus against each human gene. Next, we ranked the correlation of codon usage between viral and human genes from highest to lowest. We corrected for multiple tests using a Bonferroni correction; the significance threshold used was 7.09×10^{-9} (0.05/7,052,621 total comparisons). We obtained the highest correlations when the viral and human protein codon usage motifs were most similar.

Human tissue comparisons

We determined which proteins were expressed in each human tissue by querying each highly correlated human protein against the Human Protein Atlas [19,20]. We checked the top correlating human proteins for each virus (113 total proteins) to determine in which tissues they were most highly expressed. While many proteins were expressed in low levels throughout the body, we were most concerned with high expression areas, and only the high expression areas were compared in this study.

Results

Of the 113 viruses analyzed, we found that on average, each viral gene in 16 viruses was significantly correlated with more than 500 human proteins (Table S2). Of the remaining 97 viruses, 58 were significantly correlated with at least 100 human proteins per viral gene, and 37 were significantly correlated with at least one human gene per viral gene on average at a p-value $<7.09 \times 10^{-9}$. Only two viruses, Human papillomavirus type 90 (NC_004104) and Human gyrovirus type 1 (NC_015630) were not significantly correlated with the codon usage of at least one human gene per viral gene, on average.

The viruses listed in Table 1 have the highest Pearson r correlation values of all comparisons made, with their codon usages strongly correlating to their host codon usages (p-value$<10^{-25}$). Four of the top 10 correlations in Table 1 belong to the group of 16 viruses that strongly correlate to over 500 human proteins per viral gene on average, and the rest of them belong to the group of 58 with significant correlations with at least 100 human genes significantly correlating to each viral gene, on average. Overall, the average correlation of the 113 viruses with the top hit from each virus was 83.1%, meaning about 83% of the codon usage bias in the virus also existed in the human host protein. Each viral protein strongly correlated to an average of 303 human genes.

To demonstrate the strong correlations in codon usage bias, we plotted codon usage for several representative viral proteins compared to the human protein with the strongest correlation (Figure 1).

Finally, we analyzed the correlations of codon usage biases for human proteins expressed in tissues infected by a specific virus. With the exception of sexually transmitted diseases (STDs), tissue information was incomplete for many viruses, and further exacerbating this problem is that many human proteins expressed in a specific tissue were also expressed in many other tissues. We report all known tissue information in Table S3, and in Table 2 list representative viruses with their highest correlating protein and affected tissues.

Discussion

The high number of proteins significantly correlated with each virus suggests that humans and human-host viruses share similar codon usage biases. For example, each of the 80 Human herpesvirus 4 (HHV-4, NC_009334) genes significantly correlated with 1 to 10,012 human genes with a median of 8,290 highly correlated human genes and an average of 1,036 highly correlated human genes. HHV-4 was previously identified as having a similar codon usage bias to its host cells [21,22], which may provide insights into the efficient proliferation of HHV-4, since it can more readily utilize host tRNA machinery in the tissue types it infects. Indeed, HHV-4 (commonly known as mononucleosis or "the kissing disease") is one of the most common viruses known to infect humans, with almost 90% of adults having antibodies suggesting previous HHV-4 infection [22]. Herpesviruses overtake host translational machinery through virion host shutoff (vhs), which limits the expression of host mRNA [23], and through the degradation of host mitochondrial DNA [24], although some herpesvirus strains act differently [25]. Our data suggest that herpesvirus is able to co-opt the translational apparatus of the infected cell by closely matching codon usage biases. The virus is able to use existing tRNAs in the cell, which are not being used by the cell due to vhs.

Furthermore, viruses such as HPV-90 (NC_004104) and Human gyrovirus 1 (NC_015630) with fewer correlating proteins typically occur less frequently in human populations. Although limited data

Table 1. Here we report the top-ten codon usage bias correlations (Pearson's r values) between a virus and a human protein with their respective p-values (all under 10^{-25}), demonstrating that viruses and proteins in their host (humans) share high codon biases. Unnamed viral proteins are designated by their position numbers in the following format—Pos: start position-stop position.

Virus Accession Number	Virus Name	Virus Protein Name	Protein Accession Number	Protein Name	Correlation %	P-value
NC_009334	Human herpesvirus 4	BALF5	NP_620124.1	RHOT2	93.6	8.64E-30
NC_007605	Human herpesvirus 4 (wild type)	BALF5	NP_620124.1	RHOT2	93.5	1.36E-29
NC_000898	Human herpesvirus 6B	U90	NP_112561.2	TEX15	93.1	6.40E-29
NC_014185	Human papillomavirus 121	E1	NP_940841.1	KBTBD3	92.8	2.53E-28
NC_001716	Human herpesvirus 7	IE1	NP_001073973.2	RBM44	92.8	3.03E-28
NC_016157	Human papillomavirus 126	Pos: 817-2640	NP_940841.1	KBTBD3	92.0	6.78E-27
NC_009333	Human herpesvirus 8	ORF75	NP_002891.1	RBP3	91.8	1.47E-26
NC_010329	Human papillomavirus 88	E1	NP_940841.1	KBTBD3	90.8	4.10E-25
NC_001806	Human herpesvirus 1	UL30	NP_055778.2	SBNO2	90.8	4.15E-25
NC_014955	Human papillomavirus 132	E1	NP_940841.1	KBTBD3	90.5	9.67E-25

Table 2. A selection of viral proteins and their top correlating human proteins, along with the human protein's documented area of expression. These results show that viral codon usage biases highly correlate with the codon usage biases of human proteins that are found within tissues that the viruses are known to promote symptomatic issues.

Accession Number	Virus Name	Virus Protein	Correlating Human Protein	Protein's Expression Location
NC_004500	HPV 92	E1	MSH4	Testis
NC_022095	HPV 179	L1	HLTF	Testis
NC_014952	HPV 128	E1	TIGD4	Testis, vagina
NC_001691	HPV 50	E1	TEX15	Testis
NC_001405	HPV 18	L1	MRC2	Soft tissue, testis, endometrium
NC_001354	HPV 41	USP7	SLC12A2	Digestive tract, breast, placenta
NC_000898	HHV 6	U90	ELTD1	Gallbladder, breast, smooth muscle
NC_019023	HPV 166	E1	OTOGL	Cervix, testis
NC_009334	HHV 4	BALF5	SPTB	Epididymis
NC_010329	HPV 88	E1	RAD51AP2	Seminal Vesicle, Fallopian Tube
NC_004500	HPV 92	E1	USP9Y	Prostate

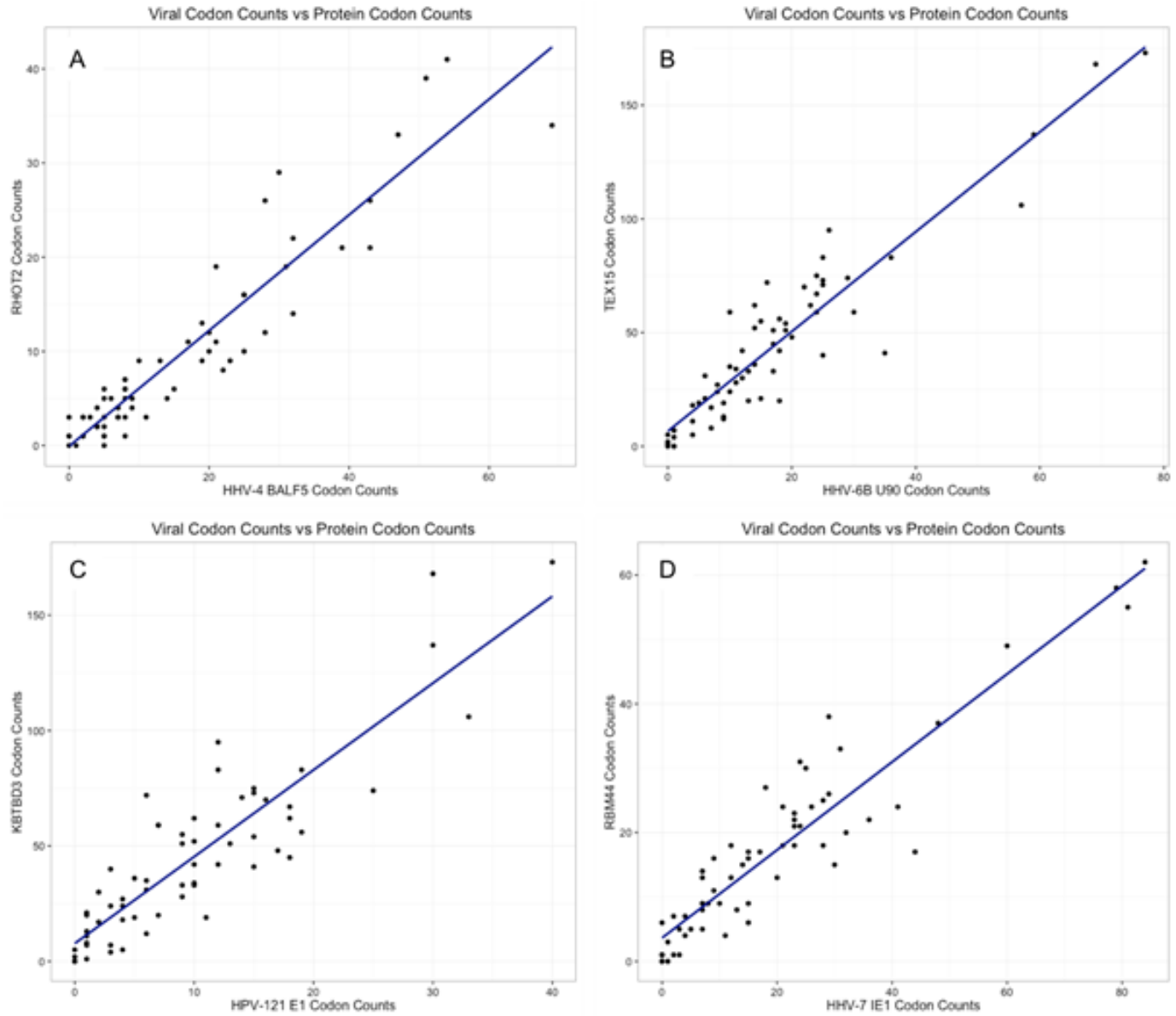

Figure 1. Codon counts. Four of the highest correlating virus-protein pairs found in Table 1 are displayed. We plotted codon counts for the viral protein (X-axis) against the human protein's codon counts (Y-axis). Each graph has 64 points, each representing a codon. Points near the top right are used at a higher rate than points near the bottom left. The line represents the result of a best-fit linear model, indicating that there is a strong correlation--as protein codon usage increases, so does the codon usage count of the respective virus. Residual plots of the linear regression were also analyzed and appear to fit the assumptions of the model. (A) displays RHOT2 vs HHV-4 (correlation of 93.6%), (B) shows TEX15 vs HHV-6B (correlation of 93.1%), (C) shows KBTBD3 vs HPV-121 (correlation of 92.8%), and (D) displays RBM44 vs HHV-7 (correlation of 92.8%). See Table 1 for more information on these pairs.

exist for the prevalence of HPV-90 in the general population, in general it presents a very low risk to the general population [26,27]. Human gyrovirus 1, which is identical to the Chicken Anemia Virus, is relatively rare and the effects of the virus still remain largely unknown, although it may affect the apoptosis pathway [28,29].

Human-host viruses appear to target tissues where the correlating human protein also has high expression. Although many viruses analyzed were not clearly annotated as infecting a particular human tissue, the viruses with documented tissue interactions were always highly correlated with a protein that was highly expressed in that tissue. For instance, HPV-128 correlates most with the human protein TIGD4, which is mainly expressed in the genitalia. In addition, other STDs were strongly correlated with proteins that were also mainly expressed in genitalia (Table 2, Table S3). We note that viruses tend to share the same codon usage biases as at least one protein that is highly expressed in the disease targeted area, further emphasizing our conclusion that viral and host codon usage biases are highly correlated.

Highly expressed genes have codon biases that utilize highly abundant tRNAs in order for optimal translational and transcriptional speed [12,13,30-33]. The Human Adenovirus E (NP_009115.2), which causes respiratory illness, has an 89.9% codon usage correlation with the NISCH gene, which is mainly expressed in the bronchus. Since NISCH is highly expressed in the tissues that the adenovirus normally infects, the virus is able to take advantage of its codon usage bias similarities with the host proteins to rapidly proliferate and infect additional hosts.

There are other possibilities for the observed shared codon usage biases. For example, co-evolution may have contributed to the appearance of such strong codon bias correlations, in which the host and the virus evolve at similar rates in order to either combat or maintain parasitic infection [34]. Since viruses have smaller genomes, they can selectively evolve more rapidly toward being similar to a preferred host.

While co-evolution and the abundance of optimal tRNAs are thought to allow greater viral spread, determining the exact cause of this correlation remains unexplored. Our extensive analysis of codon usage determined that a strong correlation in codon usage bias exists between human-host viruses and proteins expressed in the human tissues that they infect. Future research should focus on the causes of these correlations.

Authorship and contributorship

JM and PR conceived the idea. JM oversaw all aspects of the project. AH developed the comparison algorithms and ran the comparisons. CM and SW conducted literature searches and wrote sections of the paper. JM and PR were primarily responsible for editing the manuscript. PR mentored the project.

Acknowledgements

We also appreciate Mark Ebbert and Samantha Jensen who provided expert suggestions for the project flow and design.

Funding information

We appreciate the contributions of Brigham Young University and the Fulton Supercomputing Laboratory in supporting our research.

Competing interests

The authors declare that they have no competing interests.

Availability of data and material

All data are freely available from the NCBI database at ftp://ftp.ncbi.nlm.nih.gov/

References

1. Crick FH (1968) The origin of the genetic code. *J Mol Biol* 38: 367-379. [Crossref]
2. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2: 13-34. [Crossref]
3. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 24: 28-38. [Crossref]
4. Gutman GA, Hatfield GW (1989) Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A* 86: 3699-3703. [Crossref]
5. Zhang YM, Shao ZQ, Yang LT, Sun XQ, Mao YF, et al. (2013) Non-random arrangement of synonymous codons in archaea coding sequences. *Genomics* 101: 362-367. [Crossref]
6. Akashi H, Goel P, John A (2007) Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the Drosophila melanogaster subgroup. *PLoS One* 2: e1065. [Crossref]
7. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* 25: 568-579. [Crossref]
8. Xu W, Xing T, Zhao M, Yin X, Xia G, et al. (2015) Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution. *PLoS One* 10: e0131508.
9. Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* 42: 287-299. [Crossref]
10. Quax TE, Claassens NJ, Söll D, van der Oost J (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell* 59: 149-161. [Crossref]
11. Xu Y, Ma P, Shah P, Rokas A, Liu Y, et al. (2013) Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature* 495: 116-120. [Crossref]
12. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, et al. (2016) Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A* 113: E6117-6117E6125. [Crossref]
13. Chantawannakul P, Cutler RW (2008) Convergent host-parasite codon usage between honeybee and bee associated viral genomes. *J Invertebr Pathol* 98: 206-210. [Crossref]
14. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, et al. (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42: D756-D763. [Crossref]
15. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 42: D553-D559. [Crossref]
16. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 35: D5-D12. [Crossref]
17. Camiolo S, Melito S, Porceddu A (2015) New insights into the interplay between codon bias determinants in plants. *DNA Res* 22: 461-470. [Crossref]
18. Häne BG, Jäger K, Drexler HG (1993) The Pearson product-moment correlation coefficient is better suited for identification of DNA fingerprint profiles than band matching algorithms. *Electrophoresis* 14: 967-972. [Crossref]
19. Uhlén M, Björling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* 4: 1920-1932. [Crossref]
20. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, et al. (2015) Proteomics. Tissue-based map of the human proteome. Science, 347, 1260419. [Crossref]
21. Roychoudhury S, Mukherjee D (2010) A detailed comparative analysis on the overall codon usage pattern in herpesviruses. *Virus Res* 148: 31-43. [Crossref]
22. Virgin HW, Wherry EJ, Ahmed R (2009) Redefining chronic viral infection. *Cell* 138: 30-50. [Crossref]
23. Smiley JR (2004) Herpes simplex virus virion host shutoff protein: immune evasion mediated by a viral RNase? *J Virol* 78: 1063-1068. [Crossref]

24. Saffran HA, Pare JM, Corcoran JA, Weller SK, Smiley JR (2007) Herpes simplex virus eliminates host mitochondrial DNA. *EMBO Rep* 8: 188-193. [Crossref]

25. Duguay BA, Saffran HA, Ponomarev A, Duley SA, Eaton HE, et al. (2014) Elimination of mitochondrial DNA is not required for herpes simplex virus 1 replication. *J Virol* 88: 2967-2976. [Crossref]

26. Schmitt M, Depuydt C, Benoy I, Bogers J, Antoine J, et al. (2013) Prevalence and viral load of 51 genital human papillomavirus types and three subtypes. *Int J Cancer* 132: 2395-2403. [Crossref]

27. Quiroga-Garza G, Zhou H, Mody DR, Schwartz MR, Ge Y (2013) Unexpected high prevalence of HPV 90 infection in an underserved population: is it really a low-risk genotype? *Arch Pathol Lab Med* 137: 1569-1573. [Crossref]

28. Sauvage V, Cheval J, Foulongne V, Gouilh MA, Pariente K, et al. (2011) Identification of the first human gyrovirus, a virus related to chicken anemia virus. *J Virol* 85: 7948-7950. [Crossref]

29. Chaabane W, Cieślar-Pobuda A, El-Gazzah M, Jain MV, Rzeszowska-Wolny J, et al. (2014) Human-gyrovirus-Apoptin triggers mitochondrial death pathway--Nur77 is required for apoptosis triggering. *Neoplasia* 16: 679-693. [Crossref]

30. Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18: 199-209. [Crossref]

31. Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol* 46: 449-459. [Crossref]

32. Morton BR, So BG (2000) Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. *J Mol Evol* 50: 184-193. [Crossref]

33. Merkl R (2003) A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *J Mol Evol* 57: 453-466. [Crossref]

34. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, et al. (2008) Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev* 72: 457-470. [Crossref]