Peer Community In Evolutionary Biology

E5, the third oncogene of Papillomavirus

Hirohisa Kishino *based on peer reviews by* ***Leonardo de Oliveira Martins*** *and 1 anonymous reviewer*

Anouk Willemsen, Marta Félez-Sánchez, and Ignacio G. Bravo (2019) Genome plasticity in Papillomaviruses and de novo emergence of E5 oncogenes. bioRxiv, ver. 1, peer-reviewed and recommended by Peer Community in Evolutionary Biology. 10.1101/337477

Submitted: 04 June 2018, Recommended: 08 February 2019

Cite this recommendation as:
Kishino, H. (2019) E5, the third oncogene of Papillomavirus. *Peer Community in Evolutionary Biology*, 100067. 10.24072/pci.evolbiol.100067

Published: 08 February 2019

Papillomaviruses (PVs) infect almost all mammals and possibly amniotes and bony fishes. While most of them have no significant effects on the hosts, some induce physical lesions. Phylogeny of PVs consists of a few crown groups [1], among which AlphaPVs that infect primates including human have been well studied. They are associated to largely different clinical manifestations: non-oncogenic PVs causing anogenital warts, oncogenic and non-oncogenic PVs causing mucosal lesions, and non-oncogenic PVs causing cutaneous warts. The PV genome consists of a double stranded circular DNA genome, roughly organized into three parts: an early region coding for six open reading frames (ORFs: E1, E2, E4, E5, E6 and E7) involved in multiple functions including viral replication and cell transformation; a late region coding for structural proteins (L1 and L2); and a non-coding regulatory region (URR) that contains the cis-elements necessary for replication and transcription of the viral genome. The *E5*, *E6*, and *E7* are known to act as oncogenes. The E6 protein binds to the cellular p53 protein [2]. The E7 protein binds to the retinoblastoma tumor suppressor gene product, pRB [3]. However, the *E5* has been poorly studied, even though a high correlation between the type of E5 protein and the infection phenotype is observed. *E5*s, being present on the E2/L2 intergenic region in the genomes of a few polyphyletic PV lineages, are so diverged and can only be characterized by high hydrophobicity. No similar sequences have been found in the sequence database. Willemsen *et al.* [4] provide valuable evidence on the origin and evolutionary history of *E5* genes and their genomic environments. First, they tested common ancestry *vs* independent origins [5]. Because alignment can lead to biased testing toward the hypothesis of common ancestry [6], they took full account of alignment uncertainty [7] and conducted random permutation test [8]. Although the strong chemical similarity hampered decisive conclusion on the test, they could confirm that *E5* may do code proteins, and have unique evolutionary history with far different topology from the neighboring genes. Still, there is mysteries with the origin and evolution of *E5* genes. One of the largest interest may be the evolution of hydrophobicity, because it may be the main cause of variable infection phenotype. The inference has some similarity in nature with the inference of evolutionary history of G+C

contents in bacterial genomes [9]. The inference may take account of possible opportunity of convergent or parallel evolution by setting an anchor to the topologies of neighboring genes.

References:

[1] Bravo, I. G., & Alonso, Á. (2004). Mucosal human papillomaviruses encode four different E5 proteins whose chemistry and phylogeny correlate with malignant or benign growth. Journal of virology, 78, 13613-13626. doi: [10.1128/JVI.78.24.13613-13626.2004](https://dx.doi.org/10.1128/JVI.78.24.13613-13626.2004)

[2] Werness, B. A., Levine, A. J., & Howley, P. M. (1990). Association of human papillomavirus types 16 and 18 E6 proteins with p53. Science, 248, 76-79. doi: [10.1126/science.2157286](https://dx.doi.org/ 10.1126/science.2157286)

[3] Dyson, N., Howley, P. M., Munger, K., & Harlow, E. D. (1989). The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. Science, 243, 934-937. doi: [10.1126/science.2537532](https://dx.doi.org/ 10.1126/science.2537532)

[4] Willemsen, A., Félez-Sánchez, M., & Bravo, I. G. (2019). Genome plasticity in Papillomaviruses and de novo emergence of *E5* oncogenes. bioRxiv, 337477, ver. 3 peer-reviewed and recommended by PCI Evol Biol. doi: [10.1101/337477](https://dx.doi.org/ 10.1101/337477)

[5] Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. Nature, 465, 219–222. doi: [10.1038/nature09014](https://dx.doi.org/ 10.1038/nature09014)

[6] Yonezawa, T., & Hasegawa, M. (2010). Was the universal common ancestry proved?. Nature, 468, E9. doi: [10.1038/nature09482](https://dx.doi.org/ 10.1038/nature09482)

[7] Redelings, B. D., & Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. Systematic biology, 54(3), 401-418. doi: [10.1080/10635150590947041](https://dx.doi.org/10.1080/10635150590947041)

[8] de Oliveira Martins, L., & Posada, D. (2014). Testing for universal common ancestry. Systematic biology, 63(5), 838-842. doi: [10.1093/sysbio/syu041](https://dx.doi.org/ 10.1093/sysbio/syu041)

[9] Galtier, N., & Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Molecular biology and evolution, 15(7), 871-879. doi: [10.1093/oxfordjournals.molbev.a025991](https://dx.doi.org/10.1093/oxfordjournals.molbev.a025991)

Reviews

Evaluation round #2

DOI or URL of the preprint: **10.1101/337477**
Version of the preprint: 2

Authors' reply, 06 February 2019

Dear Hirohisa Kishino,

Thank you for your revision. We have verified the legends of Figures 2 and 6, and we confirm the explanations are correct. The correspondence analysis in Figure 2 has been performed on a symmetrical matrix containing

the unweighted and weighted Robinson-Foulds tree distances. The MDS in Figure 6 was performed on a matrix containing the distances in codon usage preferences for the different AlphaPV ORFS. I hope that this explanation clarified any doubts.

Kind Regards,

Anouk

Decision by Hirohisa Kishino, posted 05 February 2019

E5, the third oncogene of Papillomavirus. A recommendation of the preprint: Willemsen, A., Félez-Sánchez, M., & Bravo, I. G. (2019). Genome plasticity in Papillomaviruses and de novo emergence of E5 oncogenes. bioRxiv, 337477, ver. 3 peer-reviewed and recommended by PCI Evol Biol. doi: 10.1101/337477

MDS obtains a map based on the distance matrix. Correspondence analysis obtains a map that corresponds samples and categories. Based on these properties of the methods, I am afraid that Figures 2 and 6 were obtained not by correspondence analysis and MDS respectively and but by MDS and correspondence analysis respectively. Please confirm quickly whether the explanations of these figures are correct.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/337477

Version of the preprint: 1

Authors' reply, 24 January 2019

Download author's reply

Download tracked changes file

Decision by Hirohisa Kishino, posted 16 July 2018

Revise

Dear Anouk Willemsen,

Thank you for submitting the manuscript to PCI Evolutionary Biology. Now, we have received comments from two reviewers. each of the reviewers' comments. Both of the reviewers appreciate the work. However, Leonardo de Oliveira Martins raised methodological concerns on the UCA test, which is a core part of the manuscript, and made a constructive suggestion. Please read the comments carefully and revise the manuscript, responding to each of them.

Sincerely yours, Hiro

Reviewed by Leonardo de Oliveira Martins, 14 June 2018

The authors describe independent insertions of a non-coding stretch of DNA in the intergenic E2–L2 region of Papillomavirus (PV) genomes, with subsequent acquisition of coding capacity leading to clinically important novel proteins. The manuscript is well written, and describes concisely an important problem — de novo oncogene emergence — using an ingenious solution. At different phylogenetic scales, the authors tested explicitly if the inter–E2–L2 region (a highly variable and clinically relevant region of the circular PV genome) has a single common ancestry or appeared independently, given its low similarity and diverse composition. Within a particularly important clade (AlphaPV) they furthermore explored the genetic characteristics of the so-called E5 ORFs. Although the common ancestry hypothesis is properly addressed, I am afraid that the particular

model used may not be very convincing without a few modifications. I will describe this problem in more detail below, together with a few other suggestions.

 In [1] we give a hint on potential complications when using Bali-phy (and Bayesian models, in general) for model selection: the difficulty in achieving convergence, and the poor quality of the marginal likelihood estimation:

1. **Convergence:** For a moderate-to-high number of sequences, special attention must be paid to convergence when using Bali-phy. This is a bit different from the author's solution of running the BF test three times: what is needed is to check if, under each hypothesis, two or more independent runs achieve equilibrium (similar alignments, trees, LnL,...). It is not uncommon that even for a very long run the MCMC algorithm keeps trapped in a local optimum, given the complexity of the problem (assuming both the tree and the alignment are parameters).

2. **Marginal Likelihood:** The marginal likelihood calculated as through the geometric mean is known to be problematic, and the problem may not go away by multiple runs, etc. We wrote a follow-up on this UCA test later, describing a simpler way to test for common ancestry based on random permutations of the alignments [2]. Basically we reshuffle the columns of one of the clades and recalculate our "statistics" (difference in log-likelihoods under CA and IO hypotheses, tree length, or even average similarity, which doesn't rely on tree inference).

Therefore I would like to suggest a few options that may corroborate your conclusions and help convince readers of the independent ancestry of the inter-E2–L2 regions, at your discretion:

1. Show the phylogeny of the inter-E2–L2 regions assuming common ancestry used in the tests, from Bali-phy or even a faster method (muscle+RAxML). If the IO hypothesis is convincing, then the branch lengths leading to each cluster C1-C5 should be quite large, compared to the other branches.

2. Run convergence diagnostics for each Bali-phy analysis, to make sure the posterior distributions can be trusted. You can furthermore follow the alignment size or tree size along each sampling, and compare them to an optimal estimate (muscle+RAxML).

3. Decrease number of sequences. This may be essential in case the Bali-phy analysis is not converging — which may well be the case for more than a few dozen sequences. You may choose the four or five most dissimilar sequences within each clade.

4. Use a permutation-based test described above, from [2]. This may be faster than running Bali-phy even for for a restricted set of sequences, since you don't need to worry about convergence.

Notice that you don't need to add all the suggested analyses, but some further evidence for the independent origins hypothesis will be welcome.

 I am bit confused about the section "DNA Sequences in The inter-E2–L2 Region in AlphaPVs are Monophyletic but The E5 ORFs Therein Encoded are Not" (page 5): If E5β has an independent origin, then Cut should also be inferred as independently originated, unless they represent non-overlapping regions. Or maybe the inter-E2-L2 regions described on Table 2 exclude the E5 ORFs (the "non-coding regions" described in the discussion)? A diagram showing which regions are being included in each test, or at least a bit more info (e.g. if some sequences miss the E5 ORF, or about the non-coding regions) would help, even for the previous analysis (Table 1). Notice that this confusion may be a product of my limited knowledge of these genomes, but hopefully you can make these points clearer to other reader like me.

 Furthermore I have a few minor suggestions, that nonetheless can be easily addressed:

1. I would like to urge the authors to deposit the scripts and/or data on a publicly available repository (https://figshare.com/ or https://github.com/, for instance).

2. In general I missed some summary statistics about the sequence lengths and number of sequences on each analysis. Specially for the data sets subject to the common ancestry test, what is the average sequence length, and the equivalent alignment lengths (under each IO scenario and under CA)? This helps us having an idea about how the alignment optimisation may be influencing the homology assumptions, and is also helpful in interpreting the Bayes Factors (you may also describe the Bayes Factors normalised by the number of sites).

3. The authors may want to describe the multiple correspondence analysis in more detail — I could not see how this method is different from, e.g., an MDS plot. Furthermore on this figure (Figure 2) I would also include the concatenate tree from Figure 1, since it is the only phylogeny actually displayed in the manuscript. In theory even IO sequences can be included, since their branch lengths would denounce the disagreement with other trees, but then a distance like the weighted RF distance or the branch score distance (https://rdrr.io/cran/phangorn/man/treedist.html) should be used.

4. There is a typo on second paragraph of page 5, where you write "Common Ancestry (CO)" (should it be "CA"?). The authors might even drop the acronyms since they're not used further down the text. It seems that the acronym "MCA" is also not used and may be removed.

5. Bali-phy is not "under a maximum-likelihood framework" (page 2), it uses a Bayesian model.

References

[1] de Oliveira Martins, L. & Posada, D. Testing for Universal Common Ancestry. Systematic Biology 63, 838–842 (2014). http://www.ncbi.nlm.nih.gov/pubmed/24958930

[2] de Oliveira Martins, L. & Posada, D. Infinitely Long Branches and an Informal Test of Common Ancestry. Biology Direct 11 (1): 19. (2016) http://dx.doi.org/10.1186/s13062-016-0120-y

Reviewed by anonymous reviewer 1, 21 June 2018

This paper uses computational analysis to examine the evolution of the papillomavirus (PV) E5 ORF, which is located between the early and late region (the inter-E2-L2 region) of the PV genome. First, it provides evidence that the nucleotide sequence of the inter-E2-L2 region among the various PV types is not derived from a common ancestor. Instead, at least five independent events, one occurring for each PV clade, resulted in the insertion of this region. This implies that the E5 ORFs in the AlphaPVs (e.g., HPV16) and those of the DeltaPVs (e.g., BPV-1) are evolutionarily unrelated, consistent with the fact that the E5 proteins of HPV16 and BPV-1 share little amino acid sequence similarity except for their hydrophobicity. The authors next focused on evolution of the E5 ORFs from the AlphaPVs, which includes the HPVs. They show that while the nucleotide sequence of the inter-E2-L2 region of these PVs arose from a common ancestor, their E5 ORFs did not. Specifically, the E5 ORFs from HPVs with mucosal tropism arose separately from those with cutaneous tropism. Since the oncogenic HPVs are mucosal and not cutaneous, the independent evolution of the E5 ORF in these HPV types suggests a role for E5 in the oncogenic potential of HPVs. Finally, this paper shows the that E5 ORFs in AlphaPVs display characteristics of actual coding sequences. The authors propose that the PV E5 genes evolved by the de novo emergence of new protein-coding sequences from non-coding regions. They speculate that the independent emergence of the E5 ORFs in different HPV types occurred by random nucleotide addition and/or recombination during viral DNA synthesis to insert a noncoding sequence, followed by mutation to generate a new protein coding sequence. But, although the PV E5 genes arose independently, they all encode a small hydrophobic protein. The occurrence of multiple independent selection events for a small hydrophobic protein suggests that modulating cellular membrane proteins or the membrane environment by such a protein is important for PV fitness.

Overall, this paper provides an interesting scenario for the evolution of a diverse class of small viral transmembrane proteins and should be accepted for publication with minimal revision.

Minor corrections:

Page 10, 4th paragraph, 5th and 6th sentences should read: ”Experimentally, protein structures that have not been observed in nature have been isolated and shown to have biological activity. More specifically, Chacon et al., 2014, used genetic selection to isolate small artificial transmembrane proteins modeled after the BPV-1 E5 protein but lacking any preexisting sequences.”

Page 10, 5th paragraph, 4th sentence: replace ”rise” with ”raise”