We would like to thank the reviewers for their thorough reviews and helpful suggestions. Please find below in blue our answers to the reviewer comments. Changes in the manuscript have been marked in red.

Reviews

Reviewed by Leonardo de Oliveira Martins, 2018-06-14 06:57

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).

The reviewer is correct in raising this concern. Indeed in our initial data, we obtained very long MCMC runs, which did not necessarily converged. As we constructed an updated phylogenetic tree that includes new PV genomes (new figures 1 and S1), new IO hypotheses could be defined. Therefore we decided to redo our analysis using multiple short MCMC chains (minimum 3 chains of 100000 iterations) and combined these when convergence was reached. Convergence diagnostics were done with the bp-analyze.pl script included in the BAli-Phy software. These diagnostics have been made available on github as now indicated in the manuscript.
While rerunning our BAli-Phy analyses, we noticed that is hard to reach convergence for runs that are done at the nucleotide level. We have revised the literature and realized that none of the previously done CA tests have been performed at the nucleotide level, which raised our concerns about its performance. We suspect that the fact of dealing with only 4 states (A, T, G, and C) instead of 20 amino acids, will favor common ancestry due to highly similar sub-alignments, irrespective of the actual origin of the sequences. Our case seems even "worse" in the sense that the CA test was performed at

the nucleotide level for non-coding regions. These non-coding regions are difficult to align, and moreover, it does not seem correct to apply a substitution model to these sequences. For this reasons, we have eliminated the CA test results at the nucleotide level from the manuscript and only included the results for the E5 ORFs at the amino acid level. As a consequence, we significantly toned down the conclusions we draw in the manuscript.

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).

For the CA tests performed in the revised manuscript, we have also performed the random permutation test as suggested by the reviewer. To do this, we modified the script that Leonardo de Oliveira Martins wrote to perform this test. We calculated the Log likelihood and ML tree length statistics, as shown in supplementary figures S2 and S4. This alternative to test for common ancestry gave similar results as the BAli-Phy output for a recently submitted paper (https://www.biorxiv.org/content/early/2018/09/28/428912) where we tested whether the other PV oncogenes (E6 and E7) have a common ancestor. For E5 however, the BAli-Phy and the random permutation test results give contradictory results. As discussed in the revised manuscript. As mentioned above, we toned down the conclusions we draw in the manuscript.

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.
   We have eliminated the CA test results for the intergenic regions, so therefore we do not show this phylogeny. We did included however, the Bali-Phy phylogeny constructed on the E5 ORFs on a full and reduced data set (supplementary figure S3).
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).
   As mentioned above, we have performed convergence diagnostics with a script implemented in BAli-Phy (bp-analyze.pl). We have made the data public on github. From the online data, the best overview can be obtained by looking at the html file, where all diagnostics are summarized in one page. We calculated the optimal alignment and tree size (muscle+phyml) as a pre-analysis for the permutation test (folder /extra in github). Nevertheless, these estimates as so far

from the BAli-Phy estimates that we did not include this data in the manuscript. We did include information on the BAli-Phy alignment length and tree length in tables 1 and 2.

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.
   In the revised manuscript, we performed our analyses on the full data sets as well as on reduced data sets (now included in tables 1 an 2).
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.
   As described above, we have performed a permutation test on our data.

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.

These results would have further supported the hypothesis of *de novo* evolution of E5 in the inter-E2–L2 region. Unfortunately, for the reasons described above, we have eliminated our results of the CA test on the non-coding inter-E2–L2 region. This is thus no longer part of the revised manuscript. We did however make an effort to clearly indicate which genomes contain an E5 ORF and which do not (both written as well as graphically in figures 1 and S1).

■ 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).
   We have deposited our data and scripts https://github.com/anoukwillemsen/ONCOGENEVOL .
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).
   In the revised manuscript, we have included the information on the number sequences in each CA test, as well as the alignment length and tree length. We have also included information on the average sequence length of E5 and the inter-E2–L2 region.

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.

We have included the E5 tree in our correspondence analysis. For this we had to reduce the number of taxa from 77 to 69, as E5 is not present in all *AlphaPV* genomes. Instead of the RF distances that depend only on topology, we have followed the reviewer's suggestion and calculated the weighted RF distance that also considers the edge weight. The tree from Figure 1 was not included in this analysis as in here we only analyze PV genomes belonging to the *AlphaPV* genus, and the tree in Figure 1 contains taxa outside of this genus. We have included more details on this analysis in the Materials and Methods.

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.
   We have corrected the typo and we have made sure that we use the acronyms for CA and IO. The acronym MCA has been removed from the text.
5. Bali-phy is not "under a maximum-likelihood framework" (page 2), it uses a Bayesian model.
   Indeed this is a big mistake in the text, we have corrected this.

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, 2018-06-21 02:14

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."

This has been corrected.

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

This has been corrected.