Response to authors

I really appreciate your effort in addressing my concerns. However, I still have some that need to be clarified.

- page 5 line 184:

As the reviewer correctly deduced, we did not find homolog sequences in public databases for ORFs 5, 72, 83, 87, 94 and 107 (6 loci), thus explaining the absence of outgroups in these phylogenies. However, I'm sure that the reviewer would agree that these phylogenies are not inconsistent with the hypothesis. Obviously, they cannot! However, the rooting method is a mid-point rooting method that always places LbFV as the "outgroup" from this analysis, and one can notice that the relative distance between LbFV and the three Leptopilina species is visually very consistent among all 13 phylogenies. I think that this verbal argument brings support to our interpretation that those 6 loci derive from an LbFV ancestor (or a relative of it). In addition, the overall dataset strongly suggests that a single event led to the integration of these 13 loci (knowing that 8 of them are on the same contig in L. boulardi for instance).

For all these reason, we think that the phylogenies of all the 13 genes should be shown.

However, we agree that this was not clearly stated. We thus rewrote this part according to the reviewer's comment.

Indeed I do agree the phylogenies of the 6 loci in question are not in any way inconsistent with the authors' hypothesis. However, they are not consistent with it either. Again, with the lack of an outgroup they just do not provide evidence for a "horizontal transfer from an ancestor of the virus LbFV". I understand using midpoint rooting, however this just indicates that LbFV is very distant to the genome-insertion-event copies, and that these are closely-related. What thie ephylogenies do provide evidence for, is for a putative single origin, all these being phylogenetically very closely related. So, I suggest rephrasing as such (or similar):

"The evolutionary history of 7 genes is consistent with a horizontal transfer from an ancestor of the LbFV virus (or a virus closely related to this ancestor) to Leptopilina species (Figure 3**B-D[etc..]**). For 6 genes (ORFs 58, 78, 92, 60, 68, 85, 96), no homologs were available in public databases apart from their homologs in LbFV. However, the three copies from wasp genomes always formed a highly supported monophyletic clade."

I would argue the topology within the monophyletic clade is not very stable, having 3 with Lb as sister to (Lc + Lh), 2 with Lh sister to the other two, and one with Lc sister to the other two.

Authors could also reorder the phylogeny panels so as to group the ones that had no other homologues but LbFV.

- page 4 line 133:

We added the blast version used in the method section. The unique filter used during the blast analysis, as indicated in the method section, was based on an e-value threshold (0.01). However, LbFV hits had their e-values between 10 -5 and 10

-178 . We included this information in the text. Regarding the presence of other virus derived loci; we agree that some virally-derived genes may still rely in this genome . One could find them by performing an approach without a-priori. That was beyond the scope of this paper. We preferred to focus our attention on the exchanges that occurred between the wasps and this peculiar virus, whose biology in relation to the wasp is well known.

I understand, I assumed that is why you did not performed the searches. But I would recommend to include it in the text as a goal of the article. If not, It might leave the reader with the impression that the authors only indeed found LbFV hits and not from other viral lineages (from Nudiviridae or others), especially when some headers state things like this: "Leptopilina species captured 13 viral genes". More appropriately it should read "Leptopilina species captured 13 viral genes **from an LbFV-like virus**"

My I suggest including a phrase as such:

"**In order to identify putative events of integration from an LbFV-like virus to the wasp genomes**, we blasted the 108 proteins encoded by the behaviour-manipulating virus that infects L. boulardi (LbFV) against the Leptopilina and Ganaspis genomes (tblastn)."

- Because the length of the alignment is small, the phylogeny based in the sequencing of the PCR product (ORF96) is not very informative. Thus, several nodes are not well supported, for instance the branch of L. heterotoma and L. victoriae. If one compare the two phylogenies by taking into account only well supported clades, then they are similar. Although some specific parts of the ORF96 tree (gene tree) are not identical to the ITS tree (species tree), those parts are not supported. Thus we can say that the gene tree is not discordant with the species tree.

I would argue the topologies are generally congruent (since it is only congruent in the well-supported clades). The well-spoorted clades (going from the authors' definition of >=0.95) I see in phylogeny A are L. australis+L. clavipes, L. orientalis+(L. freyae+L .boulardi **[this clade is missing support value]**) and L. guineaensis + (L.victoriae+L. heterotoma**[not well-supported]**). The ones I see in phylogeny B are L. freyae + L. boulardi **[missing support value, I am assuming it is well supported; bipartition not well supported in phylogeny A]**, L. clavipes + L. clavipes **[same species]**, L. guineaensis + (L. clavipes + L. freyae +L. boulardi), (L.victoriae+L. heterotoma **[not well-supported in phylogeny A]**). So, I would say the topologies are generally congruent.

- page 11 line 399: The authors state “Several recent publications suggest that large, possibly full-genome insertions of symbiont into their host DNA do occur in the course of evolution, including from dsDNA viruses.”, but fail to cite the “several recent publications. Please cite these.

Sorry, my bad.

- This concatenated protein phylogeny (based on highly conserved protein set) for sure tells us the true species story. I don’t see no reason not to give it this denomination. Please correct me if I’m wrong.

Indeed, such a concatenated protein phylogeny is possibly a very good (if not the best) approximation of the species-tree, but "for sure the species tree", not. There are many approximations to a true "species-tree" (using the lax definition of any tree where several genes [protein-coding and not] are used for phylogenetic inference). The authors themselves are using yet another definition of "species-tree" in their article in figure S4, were they, in my opinion, wrongfully use the term species-tree for a phylogeny based on ITS2 sequences (single locus). So, I suggest to correct the naming of species-tree for the phylogeny based only on ITS2 and to use " a species tree was approximated".

- I think this may be useful to people not familiar with genomics and tools like BUSCO.

The thing with coverage is that, while it might give you a sense (and I mean a sense in the subjective opinion sense) about having "all" your genome sequenced, it tells you more about the quality of the base calling than of the completeness of your assembly. Completeness of your assembly (with the technology you chose for sequencing) is best estimated with k-mer analysis checking for saturation (therefore you know no new k-mers are discovered with further sequencing using the same technology) and secondarily by a BUSCO result that tells you you have all conserved genes. To my knowledge, there is no study that analyses across several eukaryotic genomes and correlates a certain minimum coverage with "genome completeness", or "sufficiency to get the whole gene set". So, I would abstain of making such a definite statement as "which is most likely sufficient to get the whole gene set".

Sincerely,

Alejandro Manzano Marín