Peer Community In
Evolutionary Biology

Open Access

APPENDIX

Editorial correspondence

Of the preprint:
Di Giovanni D, Lepetit D, Boulesteix M, Ravallec M, and Varaldi J. (2018). A behavior-manipulating virus relative as a source of adaptive genes for parasitoid wasps. *bioRxiv* 342758, ver. 5 peer-reviewed and recommended by *PCI Evol Biol*. DOI: 10.1101/342758

Peer-reviews, decisions and author's responses

Recommendation DOI: 10.24072/pci.evolbiol.100062
Recommender: Ignacio G Bravo

Based on reviews by: Alejandro Manzano-Marín and one anonymous reviewer

Revision round #2

2018-11-23
Dear Dr. Varaldi

thank you very much for having submitted the revised version of your text to PCI Evol Biol. The same two reviewers that evaluated your text in first instance have provided feedback on the resubmitted one. As you can see in the accompanying reviews, both reviewers acknowlege that you have integrated a large number of their concerns, and I agree with them. Nevertheless, also both reviewers find that some of the points raised have not been properly addressed. I share also this view in what regards to the verbal argumentation of the consistence with the main hypothesis of partial trees that do not present all terminal taxa. I think this is a view that needs be substantiated on quantitative terms rather than on verbal ones. Also, the reviewers mention some instances in the text in which the wording may suggest that you have evidence to sustain a claim for the physical existence of VLPs, that need be substantiated.

Overall, I think that the comments raised by the reviewers can be properly addressed after minor changes to the text, and I think that if you respond in detail point-to-point to each of the criticisms, a third round of peer-review should not be needed.

I sincerely thank you for supporting PCI, and look forward to read your response to the reviewers and the revised version of your text.

Sincerely

Ignacio Bravo

Preprint DOI: https://doi.org/10.1101/342758

Reviewed by Alejandro Manzano-Marín, 2018-10-05 22:56

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

Reviewed by anonymous reviewer, 2018-10-05 22:57

The authors have made several revisions based on my comments. The authors agree that a viral origin does not discount the potential status as an organelle; however, this agreement is not reflected within the paper. An example of this is in the following paragraph:

Because 323 the proteins wrapped within the VLPs have a eukaryotic origin and because 324 neither viral transcripts nor viral proteins had been identi_ed from venom 325 gland analysis, it has been claimed that VLPs do not have a viral origin [56], 326 and thus other denomination has been proposed in lieu of VLP [29]. On the 327 contrary, our data strongly suggest that the VLPs found in Leptopilina do 328 have a viral origin and derive from a massive endogenization event involving 329 a virus related to an ancestor of the behaviour manipulating virus LbFV(Fig 330 2B).

This sentence is puzzling given the fact that the authors contend in their response document: "Nowadays, VLPs are eukaryotic structures (organelles) even if some of the key genes involved in their production derive from virus genes."

Taken together, these statements are confusing. Origin of VLPs is discussed without a clear description of our current knowledge of VLPs from various Leptopilina species. Have VLPs been described from the Leptopilina species studied here? What proteins are present in VLPs and do the results in this paper have anything in common with any described VLP proteins? Even a negative result would be worth stating and discussing.

Similarly, in line 401, the authors write: "All together, our data show that VLP production is possible thanks to the domestication of 13 virally-derived genes, captured from an ancestor of LbFV."

Thus, it is clear, the authors are convinced of their idea, even though they have stepped back (superficially) by changing the title of their paper.

In this reviewer's view, the authors have not shown that the 13 virally-derived genes in the Leptopilina genomes studied are involved in VLP production. Their data demonstrate the existence of these genes in wasp genomes and their spatial and temporal expression in venom gland extracts.

As such, these results do not link LbFV genes to venom production, VLP production, or venom/VLP function. It is commendable that the team tried RNA interference experiments. However, in the absence of such results, it is advisable to wait to get the necessary evidence that will shed light on the function(s) of these interesting wasp genes/proteins. Only two sequences have any sequence similarities, but no further experimental data on these or any of the wasp LbFV proteins is available. Thus, there is a significant gap between evidence and interpretation.

Ref 56 has interesting ideas that differ from the ones proposed by the authors. It is worth stating them clearly with underlying evidence. Instead of taking an oppositional view (as in lines 320-330), a more balanced view of pertinent ideas would improve this manuscript and benefit the quality of discussions in this growing field.

Other comments:

(1) I do not understand the reluctance to show alignment of wasp ORFs with viral ORFs. This information would be informative in understanding, for example, where the primers (for expression and copy number studies) bind.

(2) If you have done experiments to check the copy number of shake and actin (used as controls for copy number), provide your evaluation of their copy number in the supplement. In the same context, provide appropriate citations showing that these genes are single copy genes in related genomes.

(3) Regarding the eukaryotic origin of LbGAP, the reference cited is incorrect. Ref. 17 is the correct reference for this. Please make sure all statements are correctly corroborated with appropriate references.

(4) The species for the Ganaspis wasps is changed in this revision. Identification of these wasps is quite difficult. So it is important to say how verification of species used was carried out. Please do this for all species. What criteria were used? Looks like G.x (line 525)—is carried over from the previous version of the paper? Please correct this.

(5) Figure S1 legend: last sentence requires a full stop at the end. Dr. Shubha Govind's name is misspelled in the acknowledgements. Please review the paper for similar errors.

Author's reply:

Dear PCI recommender,

please find below the responses to the reviewer's comments. Regarding the main issue raised by reviewer one (the interpretation of gene trees with or without virus outgroups), we followed your own suggestion and replaced our verbal argumentation by a quantitative one. We measured the mean divergence of LbV with wasps species and compared this to the divergence among wasp species for all 13 phylogenies. This relative divergence was the same for both groups (phylogenies having other viral outgroups versus phylogenies without such viral outgroups), suggesting that they do indeed have the same evolutionary history. We believe that this new argument strengthens our initial conclusion based on (among others) verbal arguments. We hope that this new version will fits your expectation.

Sincerely

Julien Varaldi

The authors have made several revisions based on my comments. The authors agree that a viral origin does not discount the potential status as an organelle; however, this agreement is not reflected within the paper. An example of this is in the following paragraph:

Because 323 the proteins wrapped within the VLPs have a eukaryotic origin and because 324 neither viral transcripts nor viral proteins had been identi_ed from venom 325 gland analysis, it has been claimed that VLPs do not have a viral origin [56], 326 and thus other denomination has been proposed in lieu of VLP [29]. On the 327 contrary, our data strongly suggest that the VLPs found in Leptopilina do 328 have a viral origin and derive from a massive endogenization event involving 329 a virus related to an ancestor of the behaviour manipulating virus LbFV(Fig 330 2B).

This sentence is puzzling given the fact that the authors contend in their response document: “Nowadays, VLPs are eukaryotic structures (organelles) even if some of the key genes involved in their production derive from virus genes.”

Taken together, these statements are confusing. Origin of VLPs is discussed without a clear description of our current knowledge of VLPs from various Leptopilina species. Have VLPs been described from the Leptopilina species studied here? What proteins are present in VLPs and do the results in this paper have anything in common with any described VLP proteins? Even a negative result would be worth stating and discussing.

Similarly, in line 401, the authors write: “All together, our data show that VLP production is possible thanks to the domestication of 13 virally-derived genes, captured from an ancestor of LbFV.”

We have modified the sentence as : “All together, our data *strongly suggest* that VLP production is possible thanks to the domestication of 13 virally-derived genes, captured from an ancestor of LbFV”

Thus, it is clear, the authors are convinced of their idea, even though they have stepped back (superficially) by changing the title of their paper.

In this reviewer's view, the authors have not shown that the 13 virally-derived genes in the Leptopilina genomes studied are involved in VLP production. Their data demonstrate the existence of these genes in wasp genomes and their spatial and temporal expression in venom gland extracts.

As such, these results do not link LbFV genes to venom production, VLP production, or venom/VLP function. It is commendable that the team tried RNA interference experiments. However, in the absence of such results, it is advisable to wait to get the necessary evidence that will shed light on the function(s) of these interesting wasp genes/proteins. Only two sequences have any sequence similarities, but no further experimental data on these or any of the wasp LbFV proteins is available. Thus, there is a significant gap between evidence and interpretation.

Ref 56 has interesting ideas that differ from the ones proposed by the authors. It is worth stating them clearly with underlying evidence. Instead of taking an oppositional view (as in lines 320-330), a more balanced view of pertinent ideas would improve this manuscript and benefit the quality of discussions in this growing field.

We do think that our data strongly suggest that VLPs do have a viral origin. That's why we propose a different scenario as the one proposed in Heavner et al. (2017) or in Poirié et al (2014) for instance. However, to give a more balanced view on this topic, we included the main arguments proposed by the authors favoring this non-viral origin hypothesis (in the discussion). But again, although we believe that our data strongly suggest that VLPs do have a viral origin, this is not contradictory to the eukaryotic origin of the proteins that are inside the VLPs (the virulence proteins). The viral genes are "only" responsible for the production of the membrane surrounding virulence proteins thus favoring their delivery to *Drosophila* immune cells.

Other comments:

(1) I do not understand the reluctance to show alignment of wasp ORFs with viral ORFs. This information would be informative in understanding, for example, where the primers (for expression and copy number studies) bind.

We included the requested alignments as supplementary figures (S2-S14).

(2) If you have done experiments to check the copy number of shake and actin (used as controls for copy number), provide your evaluation of their copy number in the supplement. In the same context, provide appropriate citations showing that these genes are single copy genes in related genomes.

We added some details in the material an methods to justify the conclusion that they are single copy genes:

Shake and actin genes were chosen as single copy genes. This was checked by looking at the blast results using each primer set (a single 100\% match was observed for both pairs of primers). Accordingly, a single band of the expected size was observed on a gel and the expected sequence was obtained after Sanger-sequencing for both loci.

(3) Regarding the eukaryotic origin of LbGAP, the reference cited is incorrect. Ref. 17 is the correct reference for this. Please make sure all statements are correctly corroborated with appropriate references.

Thank you for that. We corrected the error.

(4) The species for the Ganaspis wasps is changed in this revision. Identification of these wasps is quite difficult. So it is important to say how verification of species used was carried out. Please do this for all species. What criteria were used? Looks like G.x (line 525)—is carried over from the previous version of the paper? Please correct this.

We now mention the origin of the strains and corrected the Gx error. The identity of the Ganaspis strain has been determined by the laboratory that kindly sent us the line. The Leptopilina strains have been captured in France by our group and we used classical criteria to distinguish the two species (L. boulardi/ L. heterotoma). This is quite easy since they are the sole *Leptopilina* species in this geographical area. We do not think that this detail is necessary in the manuscript but please let us know if you think it is.

(5) Figure S1 legend: last sentence requires a full stop at the end. Dr. Shubha Govind's name is misspelled in the acknowledgements. Please review the paper for similar errors.

Thank you for that. We apologize again and corrected the error.

Reviewer 1

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 3B-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.

The major comment concerns the interpretation of the 6 "unrooted" phylogenies as opposed to the "rooted" with other viruses. To add another quantitative argument to this debate, we calculated the mean divergence of LbFV-Leptopilina relative to the mean divergence among Leptopilina species for all 13 loci. We then compared this index between groups (the 7 "rooted" versus the 6 "unrooted"). There was no difference, further suggesting that all 13 genes have the same evolutionary history. We hope that this quantitative argument, which complements other arguments (based on the similar

topologies of the phylogenies, and on the co-occurrence on the same scaffolds), will be enough to convince the reviewer. We modified this part of the text accordingly.

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)."

We modified the text according to the suggestion (with a slight modification to account for the fact that we were looking both for endogenization into wasp genomes and gene capture by the virus).

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.

We agree on that. Just a clarification: the clades without support values have low aLRT (<0.7) and thus are note reliable.

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".

We agree an modified the text accordingly (line 222).

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".

We added a reference of such a work on fish genomes. They tested the relationship between coverage and “completeness” measured as the quantity of BUSCO genes found and observed that above 15x they were done for gene content. We also have similar data on 35 hymenoptera genomes and found similar trend with slighty higher value but still below the coverage obtained in the present paper.

M. Malmstrøm, M. Matschiner, O. K. Tørresen, K. S. Jakobsen, and S. Jentoft. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Scientific Data, 4:160132, Jan 2017.

Revision round #1

2018-10-05
Dear author

thank you very much for submitting your preprint to the open PCI review process. As you can see in the accompanying files, two experts in the field have provided feedback on your submission. Both of them agree on the importance of the question and on the pertinence of the approaches used to tackle it, and I largely agree with them. Also both reviewers have invested a considerable amount of time and energy in providing a detailed report on the manuscript, with an overall very positive judgement on the methods and approaches. Notwithstanding, also both reviewers identify a number of flaws in the text that prevent recommendation in its present state, essentially centered in the , and I also largely agree with them. As you can see in their reviews, some concerns have been raised about the logical flow between results and interpretation. This is the case for instance for the support for the HGT event in the case of all genes depicted in figure 3; for the pertinence of the dn/ds values of conserved arthropod genes used to serve as reference for a set of genes only present in a subtree of these species. In other cases the questions are rather the pertinence and clarity of the figures used, as in fig 1 and fig 2. Finally, I would appreciate if you could include an assessment of the identity of the specimens actually used for the analyses, specially for the case of Ganaspis xanthopoda, as no RNA sequences were available in the screened databases.

Overall, the reviewers and myself are very supportive for recommendation upon revision. I would thus encourage you to respond to each and every point raised in the reviews. In case you think a verbose answer suffices for certain points, I would appreciate if you could make a clear case of scientific cost-benefit evaluation to justify that no novel data generation or analyses are needed.

I look very much forward to reading the revised version of the manuscript.

Faithfully

Ignacio Bravo

Preprint DOI: https://doi.org/10.1101/342758

Reviewed by Alejandro Manzano-Marín, 2018-08-17 11:06

Comments to the authors

The work presents a novel example of viral genome integration into the genome of parasitoid wasps (in this case from the *Leptopilina* genus). Contrary to previous publications, the authors were able to identify an extant dSDNA that still infects *Leptopilina boulardi* (LbFV), previously published by the authors (10.1093/gbe/evw277), as a close relative to the original viral donor. Through the use of phylogenetics and comparative genomics, they are able to provide strong evidence for a single event of integration of the viral sequences found in the analysed wasp genomes. Additionally, the authors explore the development of the venom glands and the production of the Viral-like particles (VLPs) and the expression and amplification of the virally-derived genes (VDGs) in in *L. boulardi*. Finally, extrapolating from the behavioural changes in *L. boulardi*'s egg-laying (preference to laying eggs in already parasitised larvae), the authors propose that this is a likely mechanism that could have been used by the virus to spread to other wasp lineages and could have been instrumental in the birth of the symbiotic association. I believe the article is generally OK-written (needing some restructuring and clarifications), well presented, and greatly contributes to the knowledge about how these symbiotic associations have impacted and contributed to the host biology. Most experiments and interpretations are well presented and discussed. The authors really did a well-rounded job in investigating this viral HGT to the wasp host. It deserves to be considered for publica.., sorry, recommendation after some corrections/modifications/clarifications.

Major comments

page 5 line 184: The authors state that "*The evolutionary history of the thirteen genes is consistent with an horizontal transfer from an ancestor of the virus LbFV (or a virus closely related to this ancestor) to Leptopilina species (Figure 3)*". However, Figure 3 does not precisely show that. The only phylogenies that are "consistent with a horizontal transfer from an ancestor of the virus LbFV" are those of ORF58, ORF60, ORF68, ORF78, ORF85, ORF92, and ORF96 (7 genes). For the rest of the genes, the lack of outgroups (I'm sure the authors did not found any suitable ones in the databases) does not allow the identification of the VDGs as monophyletic. The choice of LbFV as the outgroup of VDGs in those trees forces the rest of the genes to be monophyletic. Thus, if no suitable outgroup(s) to LbFV + VDGs is found, the hypothesis that the authors tried to test (VDGs monophyletic and/or sister to LbFV) is not testable. So please correct this in the results section and omit these phylogenies from the figure.

page 4 line 133: Regarding the blast results, please provide full settings (e.g. evalue threshold, percent identity) and version in the methods section. Also, please specify what you mean with "highly significant" (provide evalue or relative bitscore vs self hit or other metric). Additionally, is it correct that you only used LbFV to do the BLAST searches. How do you make sure no proteins from Nudiviridae or others that have previously been identified in other wasps are not present?

Figure 2: I find this figure is a bit confusing. First, the TEM images shown beside the *Leptopilina* seem to all be the same. I find this a bit deceiving since it gives the impression those TEMs are from each one fo the species. From reading the article they are all from *L. boulardi*. So, please remove or replace by a cartoon or other symbol. Also, the diagrams under "Wasp chromosomes", what exactly are they. Are they based on actual data or are they cartoons?. I believe this need to be corrected to make it clearer what the authors are trying to convey here.

Minor comments

Clarifications

page 2 line 38: For the phrase “However, in a number of cases”, the authors should clarify which cases by citing them. Otherwise remove the “number of cases” part.

page 2 line 41: For the sentence “The high frequency and relevance of such phenomenon has been recognized for decades for bacteria but was considered to have had a marginal impact on the evolution of metazoans” please provide citation(s) or remove.

page 3 line 85: In the sentence “However, close relatives of the donor viruses do not infect present-day wasps, nor infect their hosts.”, unless very strong evidence such as very large population surveying of these kind of viruses in a number of different species and populations is cited, authors should rephrase to something like “have not been identified, either because the ”donor” viral lineages went extinct....”.

page 4 line 122: The authors talk about repetitive sequences. How were these estimated (e.g. RepeatModeller?).

page 5 line 151: Again, what do you mean by “highly significant”, please provide numbers.

page 5 line 171: “In addition, several typical intron-containing eukaryotic genes were predicted in the vicinity of these genes (Fig. 1).”. Where exactly are these shown in figure 1?. If they are clearly not in these figure, please make a new one were it can be easily discerned.

page 6 line 210: In the phrase “The phylogeny obtained after the sequencing of the PCR products was consistent with the species-tree obtained with the ITS2 sequences (Fig. S3B).”, I just don’t see this. While the two phylogenies show some congruency, they are not perfectly congruent. For example, the clade (*L. clavipes* + *L. boulardi*) + *L. guineaensis* is indeed recovered in both phylogenies. However, the position of *L. victoriae* and *L. heteromona* are not congruent between the two phylogenies. Please rephrase this.

page 7 line 243: In “The venom gland produces the VLPs that are released in the lumen (Fig. 6) and that finally reach the reservoir where they are stored until the emergence (Fig. 5E).” I don’t believe that the VLPs reaching the reservoir can be appreciated in Fig. 5E. Please clarify or remove.

page 8 line 268: Please just clarify in a couple of words if the primers used are internal to the genes or external.

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.

page 11 line 413: Again, for “Indeed, from a function point of view, the domestication we document here is very similar to what has been described in the microgastroid complex in **B**raconidae, in **C**ampopleginae, and in **B**anchinae” add references (and capitalise).

page 12 line 430: For “[…] of the PolyDNAviruses described above) but instead proteins”, please add references.

page 14 line 508: In “We extracted the DNA of a single female abdomen using Macherey-Nagel columns, similarly to what was performed for *L. boulardi*.”, I could not find this in the text. If it is actually not in there, please cite.

page 14 line 514: Please specify the insert size of the library.

page 14 line 528: Please state the BUSCO Arthropoda database version.

page 15 line 556: Please add the version of the software used, so as to know the defaults, and/or the full list of parameters chosen.

page 15 line 561: I am uncertain about the use here of the term “species tree”. I would rather use “concatenated protein phylogeny”.

page 16 line 589: Please just specify if the primers are internal to the gene.

page 16 line 598: Where the sequences reverse-aligned with a certain software? (Pal2Nal) or an in-house script (If so please include in supplementary material).

page 18 line 667: Please specify accessed date as to know which version of database and software was used by the server.

Corrections

page 1 line 20: “[…] either because no closely related descendant infect the wasps, [...]” authors should add the possibility that the virus lineage has not yet been identified/found.

page 1 line 26: Please rephrase “Intriguingly, the contemporary [...]” to “ Intriguingly, **this** contemporary [...]” to make it clearer that you are talking about the previously referred close relative.

page 2 line 45: In “[...] leading to genetic innovation” authors could reference 10.1038/nrmicro.2017.137 (e.g. “reviewed in X”), a nice review of functional HGT from bacteria to eukaryotes.

page 3 line 81: “whereas a beta nudivirus **h**as”.

page 3 line 94: “so-called”.

page 4 line 138: In the phrase “[..] strong homologies [...]”, please correct to strong identity or similar. Sequences are either (putative) homologous or not.

page 4 line 141: Rephrase “ Two of them (**ORF** 27 and 66) are predicted” to make it more easily readable.

page 4 line 144: “In the following **section**, we will focus on the second class of genes identified by this blast analysis.”.

page 5 line 168: “by analys**ing**”.

page 5 line 169: “eas**ily** detect”.

page 5 line 177: “Taken together, th**e**se”.

page 9 line 298: Rephrase “[...] deriving from **either** a direct ancestor of LbFV **or** from a closely related one.”.

page 9 line 318: Rephrase “**, and thus o**ther denomination **h**as been proposed ***in lieu*** of VLP [26].”.

page 9 line 333: “In human**s** [...]”.

page 10 line 338: Reference 51 is weirdly located inside the parenthesis. Please check these throughout the text, as I found a couple located at weird sports in the text (e.g. ref 17).

page 17 line 621: Correct “actin**e**”.

Modifications

page 4 line 124: Either provide a citation for “ which is most likely sufficient to get the whole gene set” or just remove it. I don’t think this explanation is necessary since authors state the coverage and the BUSCO results.

page 6 line 213: This whole paragraph constitutes a conclusion, it does not represent a result. So please either move to the discussion or to a specific conclusion section.

page 6 line 228: In “showing that they are all as essential [...]” please modify to a more appropriate wording such as “**suggesting** that they are all as essential [...]”. Unless you knock them out or show otherwise they are indeed essential, “showing” is an overstatement.

page 7 line 262: I believe the second sentence in “Interestingly, among ”early” virally-derived genes, we identified a putative DNA polymerase (ORF58, see table 5). This opened the fascinating possibility that the DNA encoding those genes is amplified during this biological process.” belongs in the discussion. I suggest to leave a sentence stating the results, and the rest to be treated in the discussion.

page 8 line 287: Please add to this section the discussion sentence “ORF85 is an homologue of Ac81, a conserved protein found in all Baculoviruses” with its citation or your result from searching.

page 9 line 328: May I suggest joining this and the following paragraph. It reads nicer.

page 14 line 528: The whole phrase “For the four genomes analysed, the proportion of ”missing genes” was < 3 .5%. The statistic was even better for the three *Leptopilina* genomes (”missing genes” < 1 .9%), and the proportion of fragmented genes was also reduced compared to *Ganaspis xanthopoda* (<1 .5% for *Leptopilina* ~~**s**~~ versus 18% for *Ganaspis*).” belongs in the results section.

Tables 2, 3, 4: May I suggest to do a summary table for the main article (ORF gene and scaffold ID from the hit along with identity and alignment length). I pretty sure all of these can be summarised in a single table and the full tables can be included as a plain text (tab-separated columns) file in supplementary material.

Table 5: Again, I suggest to move this to supplementary material as a plain text file.

Figure 1: Can you please specify what are the grey “brackets” (eukaryotic genes surrounding the virally-derived genes?).

Sincerely,

Alejandro Manzano Marín

Reviewed by anonymous reviewer, 2018-08-17 11:08

This paper identifies some homologs of the behavior-modifying LbFV genes in genomic sequences of L. boulardi (where LbFV is found) and those of L. heterotoma and L. clavipes. It also addresses the possible relationship of these homologs with virulence functions of L. boulardi VLPs. The paper hypothesizes that before diversification of Figitids, LbFV captured 3 insect genes. LbFV is a descendant of this virus that then integrated into genomes of ancestral Leptopilina spp. but after its divergence from Ganaspis. The authors further claim that these integrated viral-like ORFs play a permissive role in generating the immune-suppressive Leptopilina VLPs. According to this scenario, these Leptopilina immune-suppressive VLPs are derived from erstwhile viral genes, now domesticated in wasp genomes. I have the following feedaback: Overall critique:
(1) The authors claim that the expression of the viral-like wasp genes is somehow linked to the expression of the VLP proteins but the details of this linkage are not established. No structural or functional assays establish this proposed relationship of the viral-like wasp genes with VLPs. For example, the Poirie lab has shown that RNA interference-mediated gene knockdown is possible in L. boulardi. Such an approach here would help validate if expression of the viral-like wasp genes is needed for VLP production or their function. In the absence of such functional assays, the main conclusion in the study is not supported and the authors should consider rephrasing parts of the paper, including the title.
In this context, it is important that the authors limit their interpretations for results backed by experimental data in only the wasp species for which experimental data are presented and not generalize the results to species not studied. In many places, the results are over-interpreted.
(2) Copy number experiments: It is well known that cells of the long gland portion of the venom gland cells are endopolyploid. VLP proteins are thought to be produced in these cells. I wonder if it is possible that even at the earliest stages of venom gland development, some venom gland cells undergo endopolyploidy and this affects the copy number differences observed in males and venom gland tissues. The cell type(s) in which copy number amplification is proposed to occur has not been identified. This potential difference (or change) in overall ploidy in experimental and control samples adds a wrinkle in the interpretation of the copy number data.
(3) Real time PCR experiments: The authors have previously shown that LbFV can be found in the oviduct as well as in the venom gland. It is therefore important for them show in control experiments that for the template samples used in the qPCR experiments, there is contaminating material from ovaries or related organs such as the oviduct where the viral-like wasp genes may also be expressed.
(4) Is it possible that VLPs have a viral past but the structures produced by Leptopilina wasps are not viral?

Detailed review:
- Lines 92-93: As stated above, this evidence in L. boulardi (let alone all Leptopilina wasp species) is lacking. Use of the word “permit” raises mechanistic questions for which there is no evidence or discussion.
- Line 134: Of the 17 viral proteins with significant hits in the wasp genomes: what else is known about them. A Multiple Sequence Alignment of FV genes/proteins found in the wasp genomes would highlight the dN/dS statistics they present in Figure 4 as well as introduce the predicted domains in some of these proteins where such homology exists. Are some of these domains exclusively viral or are these domains also present in eukaryotic proteins. It is important for the reader to have this information organized in a cohesive manner at the outset.
-Do viral-like genes in the Leptopilina genomes have introns? I missed this information if it is in the paper. Clarification of these points is important to understanding the hypothesis.
- Regarding proteins 27 and 66 (inhibitors of apoptosis) and 11 and 13 (the predicted methyl transferases): are there eukaryotic homologs in the sequenced Leptopilina and Ganaspis wasp genomes?
- Lines 149-150: The sentence is logically incorrect. Please restate referring to the 13 genes encoding

the proteins.
- Make a new paragraph at line 160. In the lines that follow, a new question is raised: is the depth and the GC content of scaffolds of wasp genomes with BUSCO genes and “viral-like” genes versus scaffolds with viral/bacterial genes similar or different? For the non-specialist, explain why these parameters should be similar or different in these scaffolds? This entire section is confusing and should be restructured and revised for clarity. The limitations of the results should be stated. For example, in line 177, the authors claim that their statistics “demonstrate” the presence of viral-like genes. The data are suggestive and require experimental confirmation (e.g., in situ hybridizations with appropriate probes) to actually demonstrate this. Line 180 is particularly unclear.
-Fig. 3. The data in this Figure constitute the key observation of the paper. It would be great to have experimental evidence to support the predictions of these assemblies in any one of the wasps. Otherwise, they remain predictive and should be stated as such. Molecular data showing the importance of these ORFs would validate the prediction and importance.
-Lines 252, 614 and other places. Please correct the spelling of actin.

Feedback on figures:
Fig. 1: Show in landscape . - Fig and legend 2: This Figure needs to be reworked with two clear parts (A) and (B). The legend should also be more fully developed.
-Say something about the 3 insect genes in the legend.
Sentence starting: Nowadays, all Leptopilina species bear 13 LbFV-derived… Qualify this sentence to limit only those species that were studied. Has the requirement for the 13 LbFV genes for VLP production been shown for all Leptopilina species or is this a prediction/extrapolation? If it is an extrapolation based on the PCR results of ORF 96 in the other Leptopilina species (Fig S3A), the authors should still confirm that those tested do have the rest of the expected sequences. What was the rationale for studying ORF 96 in detail?
Fig. 2: Are there any data from L. victoriae?
Fig. 2: Remove VLP panels and cite original papers that show the presence of VLPs in these species, unless these panels represent new data; if so, show them more clearly with scale bars and labeling to show VLP morphologies.
Fig. 4: Expand X axis. Words on top of the bars are not readable given the size of the graph. This should be fixed.
Fig. 5: For the light microscopy panels, make the notations and the scale bars clearer. Scale bars need to be inserted in the electron micrographs. Also, what is being observed in these micrographs is not clear. The regions need to be labeled; point to the areas with VLPs and show these areas at higher resolution and magnification to make the observation more convincing.
Fig. 7: Add L. boulardi to the legend.

Feedback on the Methods section:
- The paper states that the L. boulardi genome was deduced from a female infected with LbFV. Were the wasps used in all other experiments were infected or uninfected?
- For copy number experiments, can the primer target sequences viral and their genomic counterparts?
- Line 500, they do not state how they tested for LbFV DNA in L. heterotoma or G. xanthopoda.
-In the Methods section (lines 526-543) discuss the issue of genome sizes. However, they do not reveal if their estimates are consistent with the published work.
-How many actin genes (and how many copies of each) are predicted in the L. boulardi genome and which of these is used to control real-time PCR data? Is its expression the same in males and females?
- Explain what the single copy gene shake encodes? How do you know it is a single copy gene in these wasps?
-How did the authors determine that the RhoGAP is part of the L. boulardi VLP and is not just a protein that is part of the fluid component of the venom?

Author's reply:

New title : A behavior manipulating virus relative as a source of adaptive genes for parasitoid wasps

Dear recommender,

Please find our revised version, now entitled “A behavior manipulating virus relative as a source of adaptive genes for parasitoid wasps”. We did our best to take into account the comments of both reviewers. Reviewer 2 was not completely convinced of the link between the viral genes and the production of VLPs, in particular because we don’t have a functional test of this hypothesis. However, we believe that we provide enough evidences that allow one to conclude that those genes are indeed responsible for the production of VLPs. However, we also agree that a functional test would have been ideal to definitely prove this link. Thus, we modified the title and the text to be less affirmative regarding this link. Below are our responses to every point raised by the reviewers. We thank both reviewers for their work, that, hopefully, helped us improve the manuscript.

Best regards,
Julien Varaldi

##
Response to reviewer 1
##

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.

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.

Figure 2:

We tried to clarify this figure.

page 6 line 210: In the phrase "The phylogeny obtained after the sequencing of the PCR products was consistent with the species-tree obtained with the ITS2 sequences (Fig. S3B).", I just don't see this. While the two phylogenies show some congruency, they are not perfectly congruent. For example, the clade (*L. clavipes* + *L. boulardi*) + *L. guineaensis* is indeed recovered in both phylogenies. However, the position of *L. victoriae* and *L. heteromona* are not congruent between the two phylogenies. Please rephrase this.

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.

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.

The references are in the next two sentences.

page 15 line 561: I am uncertain about the use here of the term "species tree". I would rather use "concatenated protein phylogeny".

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.

Thank you for the reference Husnik & McClutcheon

page 10 line 338: Reference 51 is weirdly located inside the parenthesis. Please check these throughout the text, as I found a couple located at weird sports in the text (e.g. ref 17).

OK

page 4 line 124: Either provide a citation for " which is most likely sufficient to get the whole gene set" or just remove it. I don't think this explanation is necessary since authors state the coverage and the BUSCO results.

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

page 7 line 262: I believe the second sentence in "Interestingly, among "early" virally-derived genes, we identified a putative DNA polymerase (ORF58, see table 5). This opened the fascinating possibility that the DNA encoding those genes is amplified during this biological process." belongs in the discussion. I suggest to leave a sentence stating the results, and the rest to be treated in the discussion.

We decided to let it as it is, because we think this sentence renders the reading easier.

page 8 line 287: Please add to this section the discussion sentence "ORF85 is an homologue of Ac81, a conserved protein found in all Baculoviruses" with its citation or your result from searching.

OK, we included the reference. Thank you.

Other minor corrections have also been done. Thank you for the detailed review!

##########
Response to reviewer 1
##########

(1) The authors claim that the expression of the viral-like wasp genes is somehow linked to the expression of the VLP proteins but the details of this linkage are not established. No structural or functional assays establish this proposed relationship of the viral-like wasp genes with VLPs. For example, the Poirie lab has shown that RNA interference-mediated gene knockdown is possible in L. boulardi. Such an approach here would help validate if expression of the viral-like wasp genes is needed for VLP production or their function. In the absence of such functional assays, the main conclusion in the study is not supported and the authors should consider rephrasing parts of the paper, including the title.

We agree that we do not provide a functional test of our hypothesis. In fact, we tried to perform the experiment that the referee mentions (RNAi). However, we were not able to decrease the level of expression of the target gene, and thus were not able to test the hypothesis on a functional ground. There may be several reasons why this experiment was not successful. One is related to the fact that the genes targeted are expressed relatively early during pupation (starting from day 11) and that the levels of expression of those virally-derived genes are overall relatively low. This makes the experiment quite tricky, because we had to inject the dsRNA construct quite early in development (at day 11), and then had to measure the efficiency of the treatment (measure the reduction in expression level) on venom glands extracted at day 14. Unfortunately, we were not able to show a significant reduction in

the level of expression of the targeted gene. We agree that this would have been a valuable argument in favor of the proposed scenario if one can show that the level of encapsulation is reduced after this (successful) treatment. However, we provide in the paper several other solid arguments strongly suggesting that those genes are indeed responsible for the production of the VLPs in *Leptopilina* species. The arguments are (1) the genes are under strong selective pressure, as is expected for such genes, (2) they are shared by virtually all *Leptopilina* species (we will discuss this point later) as is suspected for VLP production, (3) those genes are expressed only in the tissue that is specialized in the production of the VLPs, (3) those genes are only expressed during the time period where VLPs are massively produced (4) the annotation of some of those genes suggests that they are involved in membrane metabolism. We think that all these arguments are sufficient to establish the link between those virally-derived genes and the VLP production. Finally, we argue that this scenario is not unlikely at all, if we consider the recent burst in data showing a link between virus domestication and the production of immunogenic structures in parasitic wasps (Bezier et al 2009; Volkoff et al. 2010; Pichon et al. 2015; and the very recent Burke et al. 2018). However, we took into consideration the criticism and modified the title and some key sentences in order to be less affirmative.

In this context, it is important that the authors limit their interpretations for results backed by experimental data in only the wasp species for which experimental data are presented and not generalize the results to species not studied. In many places, the results are over-interpreted.

We first studied the genome of three species belonging to the monophyletic genus Leptopilina. From this, we identified a set of thirteen genes deriving from a virus, that is shared by the three species. Those genes are absent from the outpgroup Ganaspis. From this (and additional arguments that are discussed in the text) we conclude that the genes have been acquired once by an ancestor of *Leptopilina* species. According to this hypothesis, we were able to detect the presence of the most conserved locus (ORF96) in all PCR assays involving *Leptopilina* species. This is a fairly classical reasoning in the field of evolutionary biology, ie parsimony. This scenario is much more likely (since one event may explain all the data) than alternative ones that would assume for instance multiple events explaining different outcomes in different species. However, we agree that we cannot exclude that some Leptopilina species could have lost either some of the 13 genes or the whole gene set (although this last possibility cannot concern L. boulardi, L. guineaensis, L. victoriae, L. heterotoma, L. freyae and L. clavipes that encodes for sure at least the ortholog of ORF96, as shown by the PCR assay (Fig.S3)). So from this, we argue that we can generalize the fact that all or at least most *Leptopilina* species are expected to encode the 13 virally derived genes.

Then, and because we do have limited ressources (human and financial), we studied the biology of these 13 genes only in *L. boulardi*. We previously argued that the overall dataset generated in this species strongly suggests that those genes are responsible for the production of the VLPs. Knowing that those genes are shared by all (or most) *Leptopilina* species, we extrapolated that those thirteen genes are also responsible for VLP production in other *Leptopilina* species. We agree that this is an extrapolation, but not an over-interpretation. Indeed, the dN/dS are very low for all those genes. This indicates that a strong stabilizing selection did act on those genes, at least in the genomes of *L. boulardi*, *L. heterotoma* and *L. clavipes*. This suggests that those genes have been selected for the same "function" since the divergence of these species. Based on this rationale, there is no reason to think, to our opinion, that the biological function fulfilled by those genes have changed over time.

(2) Copy number experiments: It is well known that cells of the long gland portion of the venom gland cells are endopolyploid. VLP proteins are thought to be produced in these cells. I wonder if it is possible that even at the earliest stages of venom gland development, some venom gland cells undergo endopolyploidy and this affects the copy number differences observed in males and venom gland tissues. The cell type(s) in which copy number amplification is proposed to occur has not been identified. This potential difference (or change) in overall ploidy in experimental and control samples adds a wrinkle in the interpretation of the copy number data.

We thank the referee for this information that we were not aware of. However, since we quantify the relative number of a target gene compared to a single copy gene (actine), this phenomenon cannot explain the pattern observed in figure 8.

(3) Real time PCR experiments: The authors have previously shown that LbFV can be found in the oviduct as well as in the venom gland. It is therefore important for them show in control experiments that for the template samples used in the qPCR experiments, there is contaminating material from ovaries or related organs such as the oviduct where the viral-like wasp genes may also be expressed.

All strains used in these experiments are free of LbFV. We included this information in the method section.

(4) Is it possible that VLPs have a viral past but the structures produced by Leptopilina wasps are not viral?

To my opinion, this is exactly the case! Nowadays, VLPs are eukaryotic structures (organelles) even if some of the key genes involved in their production derive from virus genes.

Detailed review:
- Lines 92-93: As stated above, this evidence in L. boulardi (let alone all Leptopilina wasp species) is lacking. Use of the word “permit” raises mechanistic questions for which there is no evidence or discussion.

In this sentence we just say that **we provide strong evidence** that these genes permit all *Leptopilina* species to produce VLPs. And I think we do.

- Line 134: Of the 17 viral proteins with significant hits in the wasp genomes: what else is known about them. A Multiple Sequence Alignment of FV genes/proteins found in the wasp genomes would highlight the dN/dS statistics they present in Figure 4 as well as introduce the predicted domains in some of these proteins where such homology exists. Are some of these domains exclusively viral or are these domains also present in eukaryotic proteins. It is important for the reader to have this information organized in a cohesive manner at the outset.

Multiple alignments used for dN/dS calculation do not include LbFV sequence since we were interested in knowing the nature of natural selection after the endogenization process.

-Do viral-like genes in the Leptopilina genomes have introns? I missed this information if it is in the paper. Clarification of these points is important to understanding the hypothesis.

We have no firm answer to that question (since this requires transcriptomic data such as RNAseq at the different stages 11, 14, 16 days, that we don't have). However, the bioinformatic predictions did not identify introns. In addition, the length of the ORF in *Leptopilina* species is highly correlated with the ORF length in the virus genome LbFV with slopes close to 1 (as indicated lines 160-162). This suggests that there is no introns. We added this conclusion in the text.

- Regarding proteins 27 and 66 (inhibitors of apoptosis) and 11 and 13 (the predicted methyl transferases): are there eukaryotic homologs in the sequenced Leptopilina and Ganaspis wasp genomes?

ORF 27, ORF6 and ORFs 11 and 13 are shared by the three *Leptopilina* genomes and also the *Ganaspsis* genome. This information is not discussed in details here since this was presented in a previous paper (Lepetit et al. 2016, GBE). However, we updated the phylogenies with additional sequences and included the corresponding phylogenies in fig. S1.

- Lines 149-150: The sentence is logically incorrect. Please restate referring to the 13 genes encoding the proteins.

Yes we corrected this. Thank you!

- Make a new paragraph at line 160. In the lines that follow, a new question is raised: is the depth and the GC content of scaffolds of wasp genomes with BUSCO genes and "viral-like" genes versus scaffolds with viral/bacterial genes similar or different? For the non-specialist, explain why these parameters should be similar or different in these scaffolds? This entire section is confusing and should be restructured and revised for clarity. The limitations of the results should be stated. For example, in line 177, the authors claim that their statistics "demonstrate" the presence of viral-like genes. The data are suggestive and require experimental confirmation (e.g., in situ hybridizations with appropriate probes) to actually demonstrate this. Line 180 is particularly unclear.

We tried to clarify this.

-Fig. 3. The data in this Figure constitute the key observation of the paper. It would

be great to have experimental evidence to support the predictions of these assemblies in any one of the wasps. Otherwise, they remain predictive and should be stated as such.
Molecular data showing the importance of these ORFs would validate the prediction and importance.

I guess that you are referring to figure 1. For sure molecular data confirming these assemblies would be interesting. However, assembly algorithms are now very efficient at reconstructing contigs so we see no reason to think that these are incorrect. In addition, the three datasets (Lb, Lh and Lc) led to the same observation, ruling out the possibility of such a technical bias.

-Lines 252, 614 and other places. Please correct the spelling of actin.

OK.

Feedback on figures:

Figs 1 & 2 : We did the requested changes. In fig. 2, we replaced TEM photos by cartoons representing VLPs instead, to be clear that this is just illustrative. We did not include *L. victoriae* in fig. 2, since in this figure we only focused on the genomes analyzed.

The rationale for studying ORF96 was that it is the most conserved gene, thus maximizing the chance of its detection. We included this information in the text.

Fig. 4: Expand X axis. Words on top of the bars are not readable given the size of the graph. This should be fixed.

We prefer to keep this scaling on x axis, because dN/dS value of 1 is of particular importance (expected value under a neutral model). As requested, we increased the size of the labels.

Fig. 5: For the light microscopy panels, make the notations and the scale bars clearer. Scale bars need to be inserted in the electron micrographs. Also, what is being observed in these micrographs is not clear. The regions need to be labeled; point to the areas with VLPs and show these areas at higher resolution and magnification to make the observation more convincing.

We redesigned the figure according to reviewer's comment.

Feedback on the Methods section:
- The paper states that the L. boulardi genome was deduced from a female infected with LbFV. Were the wasps used in all other experiments were infected or uninfected?

We included this information in the text (they are not infected).

- For copy number experiments, can the primer target sequences viral and their genomic counterparts?

The divergence between viral and wasp genes is very important (only around 30% identitiy at the protein level). Consequently there is not possibility of cross amplification. More importantly, all experiments are performed on uninfected strains.

- Line 500, they do not state how they tested for LbFV DNA in *L. heterotoma* or *G. xanthopoda*.

We tested for LbFV infection in L. heterotoma by PCR (ref 45) and by searching for LbFV reads in the genomic sequences. We included this information lines 518-519.

-In the Methods section (lines 526-543) discuss the issue of genome sizes. However, they do not reveal if their estimates are consistent with the published work.

The data are presented in table 1 with our genome estimation compared to Cytometry based estimation previously published. We included the reference to previously published data in the legend.

-How many actin genes (and how many copies of each) are predicted in the L. boulardi genome and which of these is used to control real-time PCR data? Is its expression the same in males and females?

We did not search extensively for all actin copy locus in the genome. We simply identified one of them and checked that the primers we defined did amplify a single locus. This was checked by looking at the blast results using this primer set (a single 100% match was observed for both). Accordingly, a single band of the expected size was observed on a gel. A PCR amplicon was sequenced and gave the expected unique sequence.

- Explain what the single copy gene shake encodes? How do you know it is a single copy gene in these wasps?

Shake is a gene involved in behavior in most organisms. The important point here is that it is single copy locus. Similarly to what has been performed on actin, a blast search using this primer set led to a single hit for each primer. Accordingly, a single band was observed on a gel and the sequencing of the PCR amplicon gave the expected unique sequence.

-How did the authors determine that the RhoGAP is part of the L. boulardi VLP and is not just a protein that is part of the fluid component of the venom?

We do not know precisely in fact. We rephrased the sentence to be less affirmative (line 262). What is known is that this protein is the major component of the venom produced by the

venom gland, where VLPs are extremely abundant. This is a kind of extrapolation but quite expected to my opinion, as observed in VLPs from *Venturia canescens*. Anyway, we argue that comparing the expression and amplification profile of virally-derived genes with that of the major constituent of the venom is relevant for the understanding of the system.