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.