A behavior-manipulating virus relative as a source of adaptive genes for parasitoid wasps

D. Di Giovanni[1], D. Lepetit[1], M. Boulesteix[1], M. Ravallec[2], J. Varaldi[1*]

1 Laboratoire de Biométrie et Biologie Evolutive (UMR CNRS 5558), University Lyon 1 – University of Lyon, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

2 Unité BiVi (Biologie Intégrative et Virologie des Insectes), Université Montpellier II-INRA 1231, France

* corresponding author : julien.varaldi at univ-lyon1.fr

Abstract

To circumvent host immune response, numerous hymenopteran endo-parasitoid species produce virus-like structures in their reproductive apparatus that are injected into the host together with the eggs. These viral-like structures are absolutely necessary for the reproduction of these wasps. The viral evolutionary origin of these viral-like particles has been demonstrated in only two cases and for both, the nature of the initial virus-wasp association remains unknown. This is either because no closely related descendant infects the wasps, because it has not been sampled yet, or because the virus lineage went extinct. In this paper, we provide strong evidence that the virus-like particles (VLPs) produced by endoparasitoids of *Drosophila* belonging to the genus *Leptopilina* (Hymenoptera Figitidae) have a viral origin, solving the debate on their origin. Furthermore, the ancestral donor virus still has close relatives infecting one of the wasp species, thus giving us insights on the ecological interaction that possibly allowed the domestication process. Intriguingly, this contemporary virus is both vertically and horizontally transmitted and has the particularity to manipulate the superparasitism behavior of the wasp. This raises the possibility that behavior manipulation has been instrumental in the birth of such association between wasps and viruses.

1 Introduction

Genetic information is typically passed on from generation to generation through reproduction, *ie* vertical transmission. However, at some point during the course of evolution, organisms may gain DNA from unrelated organisms, through horizontal gene transfer (HGT). Most horizontally acquired DNA is probably purged from the genomes of the population either because it did not reach the germinal cells in case of metazoan species and/or because no advantage is carried by the foreign sequence. However, natural selection may retain the foreign DNA leading ultimately to genetic innovation in the population/species [32].

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[34]. However, this view has been recently challenged due to the discovery of numerous examples of HGT in metazoans with some of them leading to genetic innovation[7]. For instance, it has been shown that some phytophagous mites and Lepidoptera deal with chemical defenses of their host plant thanks to the acquisition of a bacterial gene involved in detoxification [67]. Other very distantly related phytophagous arthropods (Aphids, mites and gall midges) independently acquired genes involved in carotenoid biosynthesis from fungal donors[48][25][15]. These carotenoid genes were previously considered as absent from animal genomes, in spite of the essential role they play on several aspects of animal biology. Based on its strong conservation in these groups, it is speculated that they have permitted genetic innovation possibly in relation to phytophagy.

Regarding the question of domestication of horizontally-transfered DNA in eukaryotes, endoparasitic wasps are of particular interest because they have repeatedly domesticated not only single genes but entire viral machineries (review in [21] and since then [10]). Endoparasitic wasps lay their eggs inside the body of other arthropods, usually other insects, ultimately killing them. Their progeny is thus exposed to the host immune system. Notably, it has been found that the ancestor of at least three monophyletic groups of endoparasitic wasps have independently domesticated a battery of viral genes allowing them to deliver either DNA encoding immuno-suppressive factors or immuno-suppressive proteins themselves [30][55]. Strikingly, in the case DNA is delivered into the host (so-called polydnaviruses, PDV), it integrates into the host hemocytes DNA and gets expressed [5][14], manipulating the host physiology and behavior, ultimately favoring the development of wasp offspring. In cases where proteins are delivered, the viral machinery permits the delivery of these virulence proteins into host immune cells, thus inhibiting the host immune response[57] [17]. In both cases, virally-derived genes are used

by the wasp to produce a vector toolset composed of capsids and/or envelopes. However, the virulence factors themselves (or the DNA encoding the virulence factors) are of eukaryotic origin, probably pre-dating the domestication event [14]. Evolution has thus repeatedly favored the domestication of kits of viral genes allowing the production of virus-like structures in the reproductive apparatus of parasitic wasps with clear functional convergence.

Although we may speculate that the intimacy of the association between the donor viruses and their parasitoid hosts has favored the exchanges, the biology of these ancestral viruses is mostly unknown. For one such domestication event (in the Campopleginae sub-family, Ichneumonidae family), the ancestral virus has not been identified at all, whereas a beta nudivirus has been identified as the donor virus for wasps belonging to the microgastroid complex of the Braconidae family. In the recently described case of a viral replacement in the lineage leading to *Venturia canescens* (Campopleginae sub-family), it has been shown that an alpha-nudivirus was the donor. However, close relatives of the donor viruses are not known to infect present-day wasps, nor to infect their hosts. One possible explanation is that the "donor" viral lineages went extinct and/or have not been sampled yet. The exact nature of the association wasp/virus that permitted such massive domestication events is thus still unclear.

In this work, we identify a new independent case of virus domestication in the genus *Leptopilina* (Family Figitidae), parasitoids of *Drosophila* larvae. We provide strong evidences that the genes of viral origin permit all *Leptopilina* wasp species to produce so-called virus-like particles (VLPs). VLPs have been known for decades in this genus([57]). They are produced in the venom gland of the wasp, are devoid of DNA but contain virulence proteins that are injected, together with the egg, into the *Drosophila* larva. They protect wasp eggs from *Drosophila* immune response ([57][16]). We show that a close relative of the ancestral donor virus is still segregating in the species *L. boulardi* and its biology has been extensively studied by our group[43][52][44][39][64]. The virus, known as LbFV, belongs to a possibly new dsDNA virus family related to Hytrosaviridae, and more distantly related to Nudiviridae and Baculoviridae[39]. The virus is vertically transmitted and manipulates the wasp behaviour by forcing infected females to lay their eggs into already parasitized larvae. This virus-induced "host-sharing" benefits to the virus since it allows its horizontal transmission to new parasitoid lineages. On the contrary, this "superparasitism" behaviour comes with a cost to wasp fitness, making it a nice example of behaviour manipulation[20]. This result suggests that symbionts such as LbFV, might have been instrumental in the birth of such association between wasps and viruses.

2 Results

We analyzed the genomic sequences of *L. boulardi*[64], *L. clavipes*[35], *L. heterotoma* (this study) and a related species in the *Ganaspis* genus (*G. brasiliensis*, this study). All *Leptopilina* species as well as *G. brasiliensis* belong to the Figitidae family and are endoparasitoids developing from various species of *Drosophila*.

The basic statistics for the assemblies used in this paper are presented in table 1. With an N50 of 2080 bp the *G. brasiliensis* assembly appeared more fragmented than those from the *Leptopilina* species whose N50 ranges from 12807 bp to 17657 bp. This reflects its two to three times larger genome size likely due to its higher content in repetitive sequences (44.92% vs. 24.02-28.82%). All four genomes were sequenced with coverage depth above 24 (between 24x and 85x), which is most likely sufficient to get the whole gene set. Accordingly, a BUSCO[59] analysis revealed that the vast majority of the 1066 single copy genes expected to be found in most arthropods are indeed present in all four assemblies (from 96.6% in *G. brasiliensis* to 99.1% in *L. boulardi*), making these assemblies suitable for HGT detection (table 1).

We inferred the relationships among the wasps under study using a set of 627 genes ubiquitous to all arthropods (see methods). As expected, the three *Leptopilina* species form a monophyletic clade with *L. heterotoma* being more closely related to *L. clavipes* than to *L.boulardi* (Fig. 1A).

We blasted the 108 proteins encoded by the behaviour-manipulating virus that infects *L. boulardi* (LbFV) against the *Leptopilina* and *Ganaspis* genomes (tblastn). Interestingly, we found that 17 viral proteins had highly significant hits in wasp genomes ($1.3 \times 10^{-178} <$ e-values $< 10^{-5}$). Among them, two classes should be distinguished. The first class is composed of four viral genes (ORFs 11, 13, 27 and 66) that have strong similarities with both *Leptopilina* and *Ganaspis* genes (Fig. S1). We previously reported that these genes have probably been acquired horizontally by the virus from an ancestral insect before the *Leptopilina* diversification ([39], Fig. S1 & 2A). Two of them (ORFs 27 and 66) are predicted to encode inhibitors of apoptosis, whereas ORFs 11 and 13 encode a putative demethylase [39]. These two last genes may derive from a single horizontal transfer followed by a subsequent gene duplication [39]. In the following section, we will focus on the second class of genes identified by this blast analysis.

2.1 *Leptopilina* species captured 13 viral genes

More surprisingly, we found clear evidence that a single massive integration of viral DNA into wasp genomes occurred before the diversification of the

Leptopilina genus and after the divergence between *Ganaspis* and *Leptopilina*. This event led to the integration of 13 viral genes into the genome of the wasps (Fig. 2B). The corresponding 13 viral proteins have highly significant hits with all *Leptopilina* species (4.10^{-4} < e-values < 1.310^{-178}, median = 10^{-33}), but not with *G. brasiliensis*. The percentages of identity between these 13 LbFV proteins and *Leptopilina* homologs ranged from 21.9 to 41.9 (table 2). All 13 loci displayed complete open reading frame (ORF) starting with a methionine and ending with a stop codon in the three wasp species, and their length was very similar to the corresponding ORF in LbFV genome (supplementary tables S1, S2 and S3; the regression slopes of ORF length in the wasp versus ORF length in LbFV were respectively 0.95, 1.02 and 0.894 for *L. boulardi*, *L. heterotoma* and *L. clavipes*; all $R^2 > 0.95$ and all p-values$< 10^{-9}$ on 11 d.f.). This suggests that those genes do not contain intron.

To define a set of expected features for typical scaffolds belonging to wasp genomes, we calculated the GC content and sequencing depth for scaffolds containing single-copy arthropod-universal BUSCO genes (Fig. S2). This is important since it allows one to distinguish genetic entities that may take part of the sample that have been sequenced. GC usually varies according to genomes, and coverage depth is directly related to the relative concentration of the DNA sequence under consideration. Except for one *L. clavipes* scaffold (scf7180005174277) encoding an homolog of ORF68, the general features (GC, sequencing depth) of wasp scaffolds sharing similarities with LbFV proteins were very similar to those calculated for the BUSCO-containing scaffolds (tables S1, S2, S3 and fig. S2). On the contrary, by analysing these statistics (GC and coverage), we could easily detect the presence of some known extra-chromosomal symbionts such as the virus LbFV in *L. boulardi* (Fig. S2A), or the bacteria *Wolbachia* in *L. heterotoma* (Fig. S2B). In addition, several typical intron-containing eukaryotic genes were predicted in the vicinity of these genes (depicted in grey in Fig. 1). Note that apart from these 13 loci specifically found in *Leptopilina* genomes, most flanking *Leptopilina* predicted proteins were also detected in the *G. brasiliensis* genome (66/72 for *L. boulardi*, 8/11 for *L. heterotoma* and 10/15 for *L.clavipes*) showing that the absence of homologs in *G. brasiliensis* genome was not the consequence of a less reliable assembly. Taken together, these observations demonstrate that the *Leptopilina* scaffolds containing viral-like genes are part of the wasp genomes. The special case of scf7180005174277 in *L. clavipes* assembly may be the consequence of recent duplications for this gene, possibly explaining its higher coverage depth.

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). Indeed, when other sequences with homology to the proteins of interest were available in public databases, the three wasp genomes always formed a highly supported monophyletic clade with LbFV as a sister group of *Leptopilina* sequences (ORFs 58, 78, 92, 60, 68, 85, 96). In addition, for the 6 remaining phylogenies (for which no homologs was available in public databases), the mid-point rooting method always led to similar topologies with LbFV as the sister group of *Leptopilina* sequences. Interestingly, it appeared from this analysis of ORF60, that before being transfered to *Leptopilina* wasps, the gene has probably been acquired by the donor virus from an ancestral bacteria (Figure 3).

The clustering of most of these loci on the same scaffold in *L. boulardi* (8 out of 13 on scaffold 159, N=75550 scaffolds, see Figure 1) strongly suggests that a single event is at the origin of the phenomenon. In addition, for a few pairs of *L. boulardi* and *L. heterotoma* scaffolds, it was possible to test for the synteny of their virally-derived genes (ORFs 92 and 107 in scaffolds 159 in Lb and IDBA_7081 in Lh, and ORFs 87 and 58 in scaffolds 2503 of Lb and IDBA_5653 in Lh). In all cases, the synteny appeared to be maintained between the two *Leptopilina* species (Fig. 1). In addition, a few flanking non-virally derived sequences were co-occuring around the same viral genes in different *Leptopilina* species (grey connections in Fig.1, see Fig. S3 for details). The overall shared organization of these genes in the three *Leptopilina* species suggests that they have been vertically inherited since a single ancestral endogenization event.

To further assess the distribution of those virally-derived genes in the diversity of *Leptopilina* wasps, we designed primers for ORF96 which is the most conserved gene. We successfully PCR amplified and sequenced the corresponding PCR product from DNA extracts obtained from all *Leptopilina* species tested (*L. guineaensis*, *L. freyae*, *L. victoriae* in addition to *L. boulardi*, *L. heterotoma* and *L. clavipes*, figure S4A). The phylogeny obtained after the sequencing of the PCR products was consistent with the species-tree obtained with the ITS2 sequences (Fig. S4B). As expected, no PCR product was obtained from *Ganaspis brasiliensis* extracts.

2.2 Virally-derived genes are under strong purifying selection in wasp genomes

In order to assess the way natural selection have acted on these virally-derived genes since their endogenization, we calculated the dN/dS ratios using alignments involving the three *Leptopilina* species. We also calculated

dNdS ratios for a set of 942 genes found in the three *Leptopilina* species and that are also shared by at least 90% of all arthropods ([59]). Those genes are thus expected to be under strong purifying selection. Accordingly, the "universal" arthropod gene set had a very low dN/dS mean value (mean=0.114, median=0.085), with a distribution skewed towards 0 (Figure 4). Interestingly, the thirteen virally-derived genes had very low and very similar dN/dS values (mean=0.215, median=0.222, min=0.125, max=0.284), suggesting that they are all as essential for the survival and/or reproduction of *Leptopilina* wasps as any "universal" arthropod gene.

2.3 Virally-derived genes are only expressed in female venom glands at the onset of VLPs production

Because *Leptopilina* wasps harbor VLPs that protect their eggs from *Drosophila* immune reaction ([57], [26]), we wondered whether the 13 virally-derived genes were in fact responsible for their production. Under this hypothesis, our prediction was that the 13 genes would be expressed only in the venom gland of females since VLPs are specifically produced in this tissue, and only when VLPs are being produced. To test this idea, we measured the expression of the 13 virally-derived genes in the venom glands, ovaries, rest of the body of *L. boulardi* females, and also in *L. boulardi* males. We followed their expression from the very beginning of the pupal stage (day 11) until the emergence of the host (day 21). During that period, the venom gland is being formed and is matured (Fig. 5). The venom gland produces the VLPs that are released in the lumen (Fig. 8) and that finally reach the reservoir where they are stored until the emergence (see the size of the reservoir in Fig. 5E).

The patterns of expression of all 13 genes fit our prediction: they are all specifically expressed in the venom glands of females but not in other tissues, nor in males (Fig. 7). Some virally-derived genes were particularly expressed at the very beginning of venom gland morphogenesis (day 11), whereas the other genes had their peak of expression at day 14, when the reservoir of the gland starts to be filled with VLPs. Two sets of genes could also be identified base on their level of expression. One set of genes had an expression between 3 and 12 times that of the actin control gene (ORFs 94, 107, 60 , 83 and 85), whereas the other genes had lower levels of expression, below 1.8 times that of the actin control (ORFs 5,72,68, 92, 87, 58, 78). ORF96 was even below the detection threshold in our assay. Finally, we also measured the expression of a wasp virulence protein, known as a major component of wasp venom, most likely wrapped within the VLPs in *Leptopilina boulardi* (the RhoGAP LbGAP [36], [17], [22]). Contrary to the 13 virally-derived genes,

this virulence protein has a eukaryotic origin ([29]). As expected, this gene is also specifically expressed in the venom gland, and transcription starts just after the 14-day peak observed for most virally-derived genes. Interestingly, among "early" virally-derived genes, we identified a putative DNA polymerase (ORF58, see table 3). This opened the fascinating possibility that the DNA encoding those genes is amplified during this biological process.

2.4 Most virally-derived genes but not the major wasp virulence factor are amplified in the venom gland

Using real-time PCR, we measured the relative DNA levels of each gene compared to an actin single copy locus. As in the transcription assay, we measured it in the venom gland, ovaries, rest of the body and in males of *L. boulardi*. We also included another single copy gene (shake) as a control. As expected the relative copy number of shake did not show any trend in time, nor differences between tissues, thus validating our assay. We observed similar "flat" patterns for ORF87, ORF58 and ORF96 although a statistically significant effect was detected at day 11 for ORFs 87 and 96. On the contrary, all other virally-derived genes were significantly amplified in the venom gland, but not in other tissues. This amplification was highly significant for most genes at day 14, were they all reached their peak of amplification. Interestingly, among the three genes that were not amplified is the putative DNA-polymerase (ORF58). This gene showed an early-transcription profile in the transcriptomic assay. The same "early-gene expression pattern" is also observed for the other non-amplified gene (ORF87). For most virally-derived genes, we observed a striking correlation between the transcription and amplification profiles (compare figs. 7 and 8). Finally, our dataset indicates that the gene encoding the major constituent of VLPs (LbGAP) is not amplified (Fig. 8).

2.5 Annotation of virally-derived genes

Out of the 13 viral genes, five had similarities with known protein domains (table 3). First, the viral protein ORF58 showed clear similarity with DNA polymerase B domain (e-value 2.310^{-20}). The domain was also detected in wasp orthologs but only for the *L. clavipes* protein. For the other four proteins, similar domains were identified in both the LbFV sequence and the wasp sequences. ORF60 bears a lecithine cholesterol acyl transferase (LCAT) domain, ORF68 contains a PIF1-like helicase, ORF78 contains an RNA-polymerase domain and ORF85 contain an Ac81 domain, a conserved protein found in all Baculoviruses [51].

3 Discussion

In this paper, we showed that all *Leptopilina* species contain a set of genes of viral origin deriving from either a direct ancestor of LbFV or from a closely related one. We describe the genomic structure of those genes in details in *L. boulardi*, *L. heterotoma* and *L. clavipes*, for which the whole genome was obtained. In addition, we were able to detect the presence of one LbFV-derived gene (ORF96) in all *Leptopilina* DNA extracts tested so far, suggesting that those virally-derived genes are shared by all *Leptopilina* species. From this analysis, we conclude that an ancestor of all *Leptopilina* species acquired a set of 13 viral genes deriving from a virus related to the behavior manipulating virus LbFV. These genes have been conserved in all *Leptopilina* species. This is very likely the consequence of a single event.

So far, all studied *Leptopilina* species are known to produce VLPs in their venom gland [57][47][26]. These spherical particles are produced at the pupal stage and are stored in the reservoir of the venom gland. During oviposition, females inject not only their egg(s) but also some VLPs into their *Drosophila* hosts. VLPs are conceptually similar to liposomes that would contain virulence proteins. VLPs then permit the wasp to address these proteins to *Drosophila* immune cells [17]. The virulence proteins delivered to the target cells then induce important morphological changes in the lamellocytes, precluding them from initiating an efficient immune reaction against the parasitoid egg [17]. Thus, the VLPs are essential for the reproduction of the wasps. Because the proteins wrapped within the VLPs have a eukaryotic origin and because neither viral transcripts nor viral proteins had been identified from venom gland analysis, it has been claimed that VLPs do not have a viral origin [56], and thus other denomination has been proposed *in lieu* of VLP [29]. On the contrary, our data strongly suggest that the VLPs found in *Leptopilina* do have a viral origin and derive from a massive endogenization event involving a virus related to an ancestor of the behaviour manipulating virus LbFV (Fig 2B).

As expected from this hypothesis, we found that the virally-derived genes are specifically expressed in the venom gland, during the first part of the pupal stage, time at which the VLPs are beginning to be produced. In addition, those genes are under strong purifying selection, as could be expected for genes involved in the production of such fitness-related structures as VLPs. Analyzing the putative biological function of the genes brings additional support in favor of this hypothesis. Although 8 out of the 13 genes had no conserved domains, three of them had functions suggesting that they could be involved in the metabolism of membrane.

The first one is ORF60 which contains a lecithine cholesterol acyl transferase

(LCAT) domain. In humans, LCAT is involved in extracellular metabolism of plasma lipoproteins, including cholesterol. LCAT esterifies the majority of free cholesterol, catalyzing translocation of fatty acid moiety of lecithin (phosphatidyl choline) to the free 3-OH group of cholesterol. It thus plays a major role in the maturation of HDL (high-density lipoprotein cholesterol) [58]. This putative biological property makes sense under our hypothesis since VLPs resemble liposomes that may be composed of highly hydrophobic compounds such as cholesterol. We may thus speculate that ORF60 plays a crucial role in the early formation of the "empty" membranes observed in the lumen of the venom gland under transmission electron microscopy (Fig. 8A-B). Interestingly, the phylogenetic reconstruction of this gene suggests that LbFV itself acquired LCAT gene from a bacterial donor species.

The second gene for which annotation could be done is ORF85. ORF85 is an homolog of Ac81, a conserved protein found in all Baculoviruses [51]. Its role has been recently deciphered in *Autographa californica* multiple nucleopolyhedrovirus (AcMNPV, [19]). During their cycle, baculoviruses first produce budded virions (BVs) and, late in infection, occlusion-derived virions (ODVs). After the initial infection, BVs are responsible for the spread of the infection from cell to cell within the infected insect. On the contrary, ODVs are only produced at the final stage of the infection. At that point nucleocapsids are retained in the nucleus where they acquire an envelope from microvesicles. They are then exported into the cytoplasm and are embedded into proteinaceous crystal matrix, thus forming occlusion bodies (OBs). The OBs are then released in the environment. OBs are absolutely necessary to initiate new insect infection through horizontal transmission. By a mutant analysis, Dong et al. [19] showed that Ac81 is necessary for the capsid envelopment and embedding within the occlusion bodies (OBs). They also showed that Ac81 contains an hydrophobic transmembrane domain that is necessary for the correct envelopment and embedding too. Interestingly, all three orthologs in *Leptopilina* sp. also contain a TM domain (Fig. S4). Our hypothesis is that the virally-derived genes found in *Leptopilina* species are responsible for the production of the VLPs, which are basically lipidic membranes. Thus we can speculate that the homolog of Ac81 in *Leptopilina* species is involved in the wrapping of proteins into the VLPs, which is observed at day 18 under electron microscopy (Fig. 8C). Interestingly, it has been found that the closest viral homolog of this protein (apart from LbFV) is a structural protein of the Hytrosaviridae GpSGHV. This is consistent with the idea that this protein is embedded into phospholipidic membranes.

The other genes containing a conserved domain suggest that the wasp has retained genes involved in DNA replication and transcription. The presence of a putative DNA polymerase (ORF58) and an helicase (ORF68) may sound

surprising if one considers that VLPs do not contain DNA, contrary to polydnaviruses. However, we observed that after the early transcription activation of the DNA polymerase (at day 11), 10 out of the 13 virall-derived genes were subsequently amplified (at day 14). This genomic amplification correlates very well with their respective expression profile which suggests that the transcriptomic regulation of these virally-derived genes is governed, at least partly, by the gene copy number in the cell. Interestingly, the DNA polymerase itself and the nearby virally-derived gene (ORF87) are not amplified, suggesting that the amplification depends on the location of the loci in wasp chromosome. It is unclear at that point whether the genomic amplification involves the production of circular or linear amplicons or concatemers, and where are located the boundaries of the amplified loci. On the contrary, the gene encoding the major constituent of the VLPs (LbGAP), which does not have a viral-origin, is not genomically amplified, although it is highly transcribed from day 14 until the emergence of the wasp. This suggests that the virally-derived DNA polymerase targets some specific sequences flanking the amplified loci. The wasp genome also encodes a virally-derived RNA polymerase (ORF78) that is likely involved in the transcription of the virally-derived genes.

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. Based on the clustering of the genes in *L. boulardi* assembly, and on the synteny conservation, we speculate that a single event led to the acquisition of the whole gene set. We can even hypothesize that a whole virus genome integrated into the chromosome of the *Leptopilina* ancestor. 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. For instance, whole genome sequencing of the brown planthopper revealed a total of 66 putative ORFs (74,730bp in total) deriving from a nudivirus genome, including 32 out of the 33 core nudiviral genes [13]. Also, it has been recently shown that an almost complete *Wolbachia* genome has been integrated into the chromosome of its host the common pillbug *Armadillidium vulgare*, with dramatic consequences on its sex determinism system[38]. After this suspected full-genome insertion of an ancestor of LbFV, we speculate that subsequent rearrangements have eliminated unnecessary genes and finally scattered, to a certain degree, the 13 remaining genes. Better genome assemblies are now necessary to gain insights on this aspect of the domestication process in the different *Leptopilina* lineages.

Our results document a novel domestication event of viruses in parasitic wasps. 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 Braconidae[4], in Campopleginae[66], and in Banchinae[3]. In all cases, it is thought that a single endogenization event led to the integration of viral DNA into wasp chromosomes, and subsequently to the evolution of a virally-derived system delivering virulence factors to host immune cells. Despite these similarities, the underlying mechanisms are different. In the braconidae *Cotesia congregata* and *Microplitis demolitor* and in the Campopleginae *Hyposoter dydimator*, the putative virally-derived genes are genomically amplified as well as the genes encoding the virulence factors[41][11][66], although different mechanisms are involved[11]. The main consequence of this amplification is the production of the DNA circles that are finally packed into the polyDNAviruses.

On the contrary in *Leptopilina boulardi*, we find that only the 13 virally-derived genes are amplified, but not the virulence gene RhoGAP. The *Leptopilina* system best resembles the VLP production observed in *Venturia canescens* in the sense that VLP do not contain DNA (contrary to PolyDNAviruses described above) but instead proteins[29]. In *Leptopilina*, the genomic amplification seems to be an original trancriptionnal mechanism occurring during the production of the VLPs membranes. To our knowledge the possibility that virally-derived gene and/or virulence factor genes are also amplified during VLP production has not been investigated in *V. canescens*.

From these examples, it is clear that the domestication of whole sets of viral genes have repeatedly occurred in endoparasitoid wasps belonging to the super-family Ichneumoinoidea, with at least two events leading to polydnavirus systems (that adress DNA circles encoding virulence factors to the host) in Braconidae and Ichneumonidae and one event leading to the evolution of a VLP system (that adress virulence proteins wrapped into a liposome-like structure to the host) in the lineage of *V. canescens* (Ichneumoinidae) [30], [55]. Actually, this last VLP domestication in *V. canescens* better corresponds to a replacement of a PDV system by a VLP system[55], showing that domestication events are frequent in this taxon. With our results, it is tempting to extend this conclusion to other distant taxons of endoparasitoids, since *Leptopilina* belongs to the family Figitidae, which diverged from Ichneumonoidea 225My ago [54].

One remaining open question for all those events, is the type of interaction the ancestral virus and its wasp did have before the domestication happened. Regarding this question, very few data are available up to now. In two probably independant cases (PDV in campopleginae such as *H. dydimator* and in banchinae such as *Glypta fumiferanae*) the ancestral virus has not been clearly identified[66][3]. On the contrary, the putative virus donors have been identified as beta-nudivirus for PDVs in braconidae[4], and as an alpha-nudivirus for VLPs found in *Venturia canescens*[55]. However, their closest

viral relatives are not infecting hymenoptera, but rather other arthropods[61]. In addition, the endogenization event is ancient, at least for Bracoviruses, which is the only case for which an estimation exists (103My, [50]), rendering difficult the inferences on the type of association that existed upon emergence of the association. It is thus unclear what type of interaction did the ancestral virus have with its host before the endogenization process.

In *Leptopilina*, we unequivocally identified an ancestor (or a close relative) of the behaviour-manipulating virus LbFV as the donor virus. First, it should be noted that in both previous cases for which the ancestor has been identified the donor virus has a large circular genome composed of a double stranded DNA. Our results again show the same pattern. Second, the previous studies repeatedly identified nudiviruses as the donor family. Here we identify a virus belonging to another, possibly new, virus family[39]. This virus is related to nudiviruses and baculoviruses, but is more closely related to the hytrosaviruses [2], which are known to induce Salivary Gland Hypertrophy in tsetse flies and house flies, although it can also remain symptom-less [1].

Finally, this is the first time that the identified virus ancestor still has extant relatives infecting one of the wasp species. Furthermore, the domestication event is more recent than the bracovirus domestication in Braconidae (103Mya, [50]), since it happened very likely after the *Ganaspis*/ *Leptopilina* divergence, which occurred around 73Mya[9]. Although this is still a large upper bound value, using this biological system may help us infer about the nature of the initial virus/wasp association. From our previous work on the interaction between LbFV and its host *Leptopilina boulardi*, we know that LbFV is vertically transmitted and replicate in cells of the oviduct[65]. This result suggests that physical proximity with the germ line may have facilitated the initial endogenization event, thus allowing the initiation of the domestication process. The identification of a contemporary virus still infecting the wasp also opens the way for addressing experimentally the mechanisms by which the virus could integrate into wasp chromosomes. Finally, LbFV is responsible for a behavior manipulation in *L. boulardi*: it forces females to superparasitize, which allows its horizontal transmission to other wasps[63]. This raises the fascinating possibility that the ancestral donor virus also manipulated the behavior of the wasp. To clarify this issue, the sampling of relatives of LbFV will be essential, to be able to reconstruct the ancestral state for the lineage that actually gave rise to such genetic innovation.

4 Methods

4.1 Wasp rearing

L. boulardi, *L. heterotoma* and *G. brasiliensis* were reared on *D. melanogaster* as host (StFoy strain) in a climatic chamber (25C 60% humidity, 12/12 LD). *Drosophila* were fed with a standard medium [18]. All experiments on *L. boulardi* were performed on a strain uninfected with the behaviour-manipulating virus (NSref).

4.2 Wasp genome sequences and annotation

We previously reported the genome of *Leptopilina boulardi*, strain Sienna (accession number : PQAT00000000) which has been obtained from the sequencing of a single female[64]. Although this female was infected by LbFV, the draft genome does not contain contigs belonging to the virus genome since we removed them by comparison to the published virus genome sequence[39]. The assembly was performed using IDBA_ud [53] followed by a scaffolding step with assembled RNAseq reads using the software L_RNA_scaffolder [69].

We sequenced the genomes of the related *L. heterotoma* (Gotheron strain), and the more distantly related *G. brasiliensis* (Va strain). *L. heterotoma* is refractory to infection by LbFV[52] and no reads mapping to LbFV genome has been found neither in *L. heterotoma* nor in *G. brasiliensis* datasets. We extracted the DNA of a single female abdomen using Macherey-Nagel columns, similarly to what was performed for *L. boulardi*[64]. The DNAs were then used to prepare paired-end Illumina libraries using standard protocols (TruSeq PE Cluster v3, TruSeq SBS 200 cycles v3, TruSeq Multiplex Primer). The libraries were then sequenced on a Hiseq2500 (for L.h, 2 x 100bp, insert size = 418bp) or Hiseq3000 (for G.x, 2 x 150bp, insert size = 438bp) machine on the Genotoul sequencing platform.

Similarly to what was done for *L. boulardi*, the drafts of *L.heterotoma* and *G.brasiliensis* were obtained after assembling genomic DNA reads with IDBA_ud [53]. For *L. heterotoma* assembly, this was followed by scaffolding using publicly available assembled RNAseq reads[22] by running the software L_RNA_scaffolder[69]. This RNA-seq scaffolding step was not performed for *G. brasiliensis* because no RNAseq reads were available for this species in public databases.

The genome of an asexual strain of *L. clavipes* (strain GBW) which is not infected by LbFV was obtained and is described in [35] (accession PRJNA84205). To have comparable assembly strategies, we included an additional RNA scaffolding step using publicly available sequences ([46]).

In order to test the completeness of the drafts generated, we ran the BUSCO pipeline (version 2.0) that looks for the presence of 1066 ubiquitous genes shared by at least 90% of all arthropods ([59]).

The genome sizes were estimated using several methods. First of all, we simply divided the total number of bases mapped to the draft by the mean coverage observed on scaffolds containing complete BUSCO genes. Those scaffolds are expected to contain non repeated nuclear DNA and their coverage is a valuable estimate of the coverage for any nuclear locus. Second, after filtering out adapters containing reads with Skewer version 0.2.2[33], removing reads duplicates with FastUniq version 1.1[68], filtering out reads mapping to mitochondrial contigs with Bowtie 2 version 2.3.4.1[37] and samtools version 1.8[40], removing contaminant reads (from viruses, prokaryotes and microbial eukaryotes) with Kaiju 1.6.2 used with the NR+euk 2018-02-23 database[45], k-mers frequencies were established from the remaining reads for each species using Jellyfish 2.2.9[42] and k = 21 (default value). From these 21-mers distributions genome size was estimated with findGSE[60] used with default parameters. These estimates were then used to run DNAPipeTE version 1.3[24] (2 samples per run, 0.1X coverage per sample) in order to assess the repetitive fraction of the genomes. Finally, independant estimates from flow cytometry experiments were obtained for *L. boulardi*, *L. heterotoma* and *G. brasiliensis* from [23] and for *L. clavipes* from [35].

We predicted genes in wasp sequences using the software augustus 3.2.3 [31], with training parameters obtained from the BUSCO outputs.

4.3 Homology search

In order to identify homologies between viral proteins and wasp DNA, we used a simple tblastn (v. 2.6.0) approach with viral proteins as query and each wasp genome as database. Default parameters were used except that an evalue threshold of 0.01 was chosen.

4.4 Phylogenies

4.4.1 Species-tree

Based on 627 "universal arthropod" genes identified by the BUSCO pipeline [59], a species tree was constructed for *L. heterotoma*, *L. boulardi*, *L. clavipes* and *G. brasiliensis*, using *Apis mellifera* as outgroup. The protein sequences were aligned using the bioconductor msa package[6]. Individual alignments were concatenated and a phylogenetic reconstruction was then performed using PhyML (parameters: -d aa -m LG -b -4 -v e -c 4 -a e -f m)[27]. In total,

290428 variable sites were found and the branch supports were computed using approximate likelihood ratio test (aLRT). We also constructed a tree for 10 *Leptopilina* species and *G. brasiliensis* using publicly available sequences of Internal transcribed spacer 2 (ITS2). Alignment was performed with muscle and a phylogeny was obtained with PhyML (parameters: -d nt -m GTR -b -4 -v 0.0 -c 4 -a e -f e). In total, 399 variable sites were used and the tree was rooted using mid-point rooting method.

4.4.2 Gene-tree

We searched orthologs of viral proteins of interest in other organisms by blasting (blastp) them against nr (downloaded on october 2017) with an evalue threshold of 0.01. After retrieving the sequences, we selected one sequence per species and added them to the proteins identified in *Leptopilina* genomes. The sequences were then aligned using muscle algorithm v3.8.31. Because the proteins included in the alignment diverged considerably, we selected blocks of conserved sites using the gblocks algorithm parametrized with less stringent options (allowing smaller final blocks, gaps within final blocks and less strict flanking positions, [12]). Phylogenetic reconstruction was then performed using PhyML (parameters: -d aa -m LG -b -4 -v e -c 4 -a e -f m). The branch supports were computed using approximate likelihood ratio test (aLRT). The accession numbers of the sequences used in the phylogenies are reported in table S4.

4.5 PCR amplification of ORF96

Based on the sequences of *L. boulardi*, *L. heterotoma* and *L. clavipes*, we designed primers for the orthologs of LbFVORF96. The primer sequences are ATTGGTGAAATTCAATCGTC and TCATTCATTCGCAATAATTGTG. They amplified a 411bp internal fragment of the coding sequence. PCR reaction was performed in a 25uL volume containing 0.2uM primers, 0.2mM dNTPs, 1mM MgCl2 and 0.5U of Taq DNA polymerase with the following cycling conditions : 95 °C 30", 54 °C 30", 72 °C 60" (33 cycles).

4.6 dN/dS calculation

The coding sequences of "universal arthropod" BUSCO genes identified in the three *Leptopilina* species were extracted and, using the msa and seqinr R package, were reverse-aligned using the protein alignments as a guide (reverse.align function of the seqinr package). dN/dS ratios were then estimated using the kaks function of the seqinr R package. The method

implemented in this package is noted LWL85 in [62]. A similar procedure was performed for the 13 virally-derived genes found in the genomes of the three *Leptopilina* species.

4.7 Expression in the venom gland and other tissues

We studied the expression of genes during the pupal stage of *L. boulardi*, at days 11, 14, 16, 18 and 21. The wasp strain used is not infected by the behaviour-manipulating virus LbFV. 11 days corresponds to the beginining of the pupal stage, whereas 21 days corresponds to the emergence time. Wasps were gently extirpated from the *Drosophila* puparium, and venom gland, ovaries, rest of the body of *L. boulardi* females was dissected in a droplet of PBS + 0.01% tween and deposited in the RLT+B-mercaptoethanol buffer of the Qiagen RNAeasy extraction kit. Males were also prepared as a control, in a similar way. The tissues extracted from twenty individuals were then pooled together and tissues were disrupted in a Qiagen homogenizer (3 minutes 25Hz). Two biological replicates were performed for each condition, except for day 11 where only one sample was obtained. cDNAs were synthetized using the SuperscriptIII kit (ThermoFisher). Real-time PCR assays were then performed with SYBR green (ssoadvanced universal sybr green supermix, Biorad) using standard procedures on a Biorad CFX-96 machine. We quantified the number of copies of each target cDNA using a serial dilution standards. Because we obtained only tiny quantities of RNA from this experiment (because of the very small size of the tissues dissected), we were not able to test numerous genes. We thus choose to use only one control gene (actin gene). As a counterpart, we were able to test all thirteen virally-derived genes and the RhoGAP gene. The primer sequences are given in table S5.

4.8 Genomic Amplification

Using a similar assay, we extracted the DNA of *L. boulardi*, at days 11, 14, 16, 18 and 21, using an uninfected strain (no LbFV present). The genomic DNA of 15 pooled individuals was extracted using the Nucleospin tissue Macherey-Nagel kit following provider's instructions. Three biological replicates per condition was done. Real-time PCR assays were then performed with SYBR green using standard procedures on a Biorad CFX-96 machine. We quantified the number of copies of each target genes using a serial dilution standards. The primer sequences are given in table S1. For an unknown reason, the amplification with DNA extracted from ovaries was particularly difficult, in particular when the ovaries were mature (at day 21). We thus had to remove this tissue from the statistical analysis because Cqs were too high to be

reliable. For the same reason, most data for ovaries at day 21 were removed from figure 8. The primer sequences are given in table S5.

4.9 Statistical analysis

For both the transcriptomic and genomic analysis, we calculated the absolute copy number of each gene of interest and divided it by the absolute copy number of the actin control gene. This ratio was then analyzed in an anova framework with time, tissue and time:tissue interation as factors. The effects were tested by likelihood ratio tests (LRT) of full model versus reduced one. Contrasts between tissues were also calculated at each time point (corresponding to the star in figures 7 and 8). Residuals of the models were judged as unstructured and had an overall normal distribution.

4.10 Morphogenesis and electron microscopy of the venom gland

To follow the morphogenesis of the venom gland, we dissected *L. boulardi* pupae at days 11, 14, 16, 18 and 21, in a similar design used for transcriptomics. Wasps were gently extirpated from the *Drosophila* puparium, and the venom gland of females was dissected in a droplet of PBS + 0.01% tween. Venom glands were either directly mounted on a glass slide for further examination under a light microscope or transfered into a solution of 2% glutaraldehyde in PBS for further examination under the Transmission Electron Microscope (TEM). For TEM, the tissues were then post fixed 1 hour in 2% osmium tetroxide in the same buffer, thoroughly rinced in distilled water, stained "en bloc" with a 5% aqueous uranyl acetate solution, dehydrated in a series of graded ethanol and embedded in Epon's medium. Ultrathin sections were cut on a LKB ultratome and double stained in Uranyless and lead citrate. Samples were examined with a Jeol 1200 Ex transmission microscope at 80kV. Images were taken with an Quemesa 11 megapixel Olympus camera and analyzed with ImageJ software (https://imagej.nih.gov/ij/).

4.11 Annotation of viral genes

We searched for the presence of conserved domains in the 13 LbFV proteins horizontally transfered to *Leptopilina* species using the hmmer webserver (https://www.ebi.ac.uk/Tools/hmmer/) accessed the 5 of may 2018.

5 Acknowledgment

This work was supported by a grant from the Agence Nationale de la Recherche (ANR) to JV (11-JSV7-0011 Viromics). The bio-informatic work was performed using the computing facilities of the CC LBBE/PRABI. We thank the labex Ecofect for financial support for the internship of DDG. We thank Shubhah Govind for kindly providing the *Ganaspis brasiliensis* strain and Ken Kraaijeveld for providing access to raw Illumina reads of *L. clavipes*. We thank PCI Evol Biol reviewers for helpful comments.

6 Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article.

References

[1] A. M. M. Abd-Alla, F. Cousserans, A. G. Parker, J. A. Jehle, N. J. Parker, J. M. Vlak, A. S. Robinson, and M. Bergoin. Genome analysis of a Glossina pallidipes salivary gland hypertrophy virus reveals a novel, large, double-stranded circular DNA virus. *Journal of virology* 82.9 (May 2008), 4595–4611.

[2] A. M. M. Abd-Alla, J. M. Vlak, M. Bergoin, J. E. Maruniak, A. Parker, J. P. Burand, J. A. Jehle, D. G. Boucias, and H. S. G. of the ICTV. Hytrosaviridae: a proposal for classification and nomenclature of a new insect virus family. *Archives of virology* 154.6 (2009), 909–918.

[3] C. Béliveau, A. Cohen, D. Stewart, G. Periquet, A. Djoumad, L. Kuhn, D. Stoltz, B. Boyle, A.-N. Volkoff, E. A. Herniou, J.-M. Drezen, and M. Cusson. Genomic and Proteomic Analyses Indicate that Banchine and Campoplegine Polydnaviruses Have Similar, if Not Identical, Viral Ancestors. *Journal of virology* 89.17 (Aug. 2015), 8909–8921.

[4] A. Bezier, M. Annaheim, J. Herbiniere, C. Wetterwald, G. Gyapay, S. Bernard-Samain, P. Wincker, I. Roditi, M. Heller, M. Belghazi, R. Pfister-Wilhem, G. Periquet, C. Dupuy, E. Huguet, A.-N. Volkoff, B. Lanzrein, and J.-M. Drezen. Polydnaviruses of Braconid Wasps Derive from an Ancestral Nudivirus. *Science (New York, NY)* 323.5916 (Feb. 2009), 926–930.

[5] K. Bitra, S. Zhang, and M. R. Strand. Transcriptomic profiling of Microplitis demolitor bracovirus reveals host, tissue and stage-specific patterns of activity. *Journal of General Virology* 92.9 (Aug. 2011), 2060–2071.

[6] U. Bodenhofer, E. Bonatesta, C. Horejš-Kainrath, and S. Hochreiter. msa: an R package for multiple sequence alignment. *Bioinformatics (Oxford, England)* 22 (Aug. 2015), btv494.

[7] L. Boto. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings Of The Royal Society B-Biological Sciences* 281.1777 (Jan. 2014), 20132450.

[8] M. L. Buffington, S. G. Brady, and S. I. Morita. Divergence estimates and early evolutionary history of Figitidae (Hymenoptera: Cynipoidea). *Systematic Entomology* 37 (2012), 287–304.

[9] M. L. Buffington, J. Nylander, and J. M. Heraty. The phylogeny and evolution of Figitidae (Hymenoptera: Cynipoidea) - Buffington - 2007 - Cladistics - Wiley Online Library. *Cladistics* (2007).

[10] G. R. Burke, T. J. Simmonds, B. J. SHARANOWSKI, and S. M. Geib. Rapid Viral Symbiogenesis via Changes in Parasitoid Wasp Genome Architecture. *Molecular biology and evolution* 15.2 (July 2018), 137.

[11] G. R. Burke, T. J. Simmonds, S. A. Thomas, and M. R. Strand. Microplitis demolitor Bracovirus Proviral Loci and Clustered Replication Genes Exhibit Distinct DNA Amplification Patterns during Replication. *Journal of virology* 89.18 (Sept. 2015), 9511–9523.

[12] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* 17.4 (Apr. 2000), 540–552.

[13] R.-L. Cheng, Y. Xi, Y.-H. Lou, Z. Wang, J.-Y. Xu, H.-J. Xu, and C.-X. Zhang. Brown planthopper nudivirus DNA integrated in its host genome. *Journal of virology* 88.10 (May 2014), 5310–5318.

[14] G. Chevignon, J. Theze, S. Cambier, J. Poulain, C. Da Silva, A. Bézier, K. Musset, S. J. M. Moreau, J.-M. Drezen, and E. Huguet. Functional annotation of Cotesia congregata bracovirus: identification of viral genes expressed in parasitized host immune tissues. *Journal of virology* 88.16 (Aug. 2014), 8795–8812.

[15] C. Cobbs, J. Heath, J. O. Stireman, and P. Abbot. Carotenoids in unexpected places: gall midges, lateral gene transfer, and carotenoid biosynthesis in animals. *Molecular phylogenetics and evolution* 68.2 (Aug. 2013), 221–228.

[16] D. Colinet, A. Schmitz, D. Cazes, J.-L. Gatti, and M. Poirié. The origin of intraspecific variation of virulence in an eukaryotic immune suppressive parasite. *PLoS pathogens* 6.11 (2010), e1001206.

[17] D. Colinet, A. Schmitz, D. Depoix, D. Crochard, and M. Poirié. Convergent Use of RhoGAP Toxins by Eukaryotic Parasites and Bacterial Pathogens. *PLoS pathogens* 3.12 (2007), e203.

[18] J. R. David. A new medium for rearing Drosophila in axenic conditions. *Drosophila Info Service* 36 (Apr. 1962), 128.

[19] F. Dong, J. Wang, R. Deng, and X. Wang. Autographa californica multiple nucleopolyhedrovirus gene ac81 is required for nucleocapsid envelopment. *Virus Research* 221 (Aug. 2016), 47–57.

[20] S. Gandon, A. Rivero, and J. Varaldi. Superparasitism evolution: Adaptation or manipulation? *The American naturalist* 167 (2006), E1–E22.

[21] J. Gauthier, J.-M. Drezen, and E. A. Herniou. The recurrent domestication of viruses: major evolutionary transitions in parasitic wasps. *Parasitology* (May 2017), 1–13.

[22] J. Goecks, N. T. Mortimer, J. A. Mobley, G. J. Bowersock, J. Taylor, and T. A. Schlenke. Integrative Approach Reveals Composition of Endoparasitoid Wasp Venoms. *PLoS ONE* 8.5 (May 2013), e64125.

[23] V. E. Gokhman, J. S. Johnston, C. Small, R. Rajwani, S. J. Hanrahan, and S. Govind. Genomic and karyotypic variation in Drosophila parasitoids (Hymenoptera, Cynipoidea, Figitidae). *Comparative Cytogenetics* 5.3 (Dec. 2010), 211–221.

[24] C. Goubert, L. Modolo, C. Vieira, C. ValienteMoro, P. Mavingui, and M. Boulesteix. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome biology and evolution* 7.4 (Mar. 2015), 1192–1205.

[25] M. Grbić et al. The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature* 479.7374 (Nov. 2011), 487–492.

[26] G. Gueguen, R. Rajwani, I. Paddibhatla, J. Morales, and S. Govind. VLPs of Leptopilina boulardi share biogenesis and overall stellate morphology with VLPs of the heterotoma clade. *Virus Research* 160.1-2 (Sept. 2011), 7–7.

[27] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59.3 (May 2010), 307–321.

[28] L. Guy, J. Roat Kultima, and S. G. E. Andersson. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics (Oxford, England)* 26.18 (Sept. 2010), 2334–2335.

[29] M. E. Heavner, J. Ramroop, G. Gueguen, G. Ramrattan, G. Dolios, M. Scarpati, J. Kwiat, S. Bhattacharya, R. Wang, S. Singh, and S. Govind. Novel Organelles with Elements of Bacterial and Eukaryotic Secretion Systems Weaponize Parasites of Drosophila. *Current biology : CB* 27.18 (Sept. 2017), 2869–2877.e6.

[30] E. A. Herniou, E. Huguet, J. Theze, A. Bézier, G. Periquet, and J.-M. Drezen. When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1626 (Sept. 2013), 20130051.

[31] K. J. Hoff and M. Stanke. WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic acids research* 41.Web Server issue (July 2013), W123–8.

[32] F. Husnik and J. P. McCutcheon. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature reviews Microbiology* 16.2 (Nov. 2017), 67–79.

[33] H. Jiang, R. Lei, S.-W. Ding, and S. Zhu. Skewer - a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics* 15 (2014), 182.

[34] P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9.8 (Aug. 2008), 605–618.

[35] K. Kraaijeveld, Y. Anvar, J. Frank, A. Schmitz, J. Bast, J. Wilbrandt, M. Petersen, T. Ziesmann, O. Niehuis, P. d. Knijff, J. T. d. Dunnen, and J. Ellers. Decay of sexual trait genes in an asexual parasitoid wasp. *Genome biology and evolution* (Nov. 2016), evw273.

[36] C. Labrosse, P. Eslin, G. Doury, J. M. Drezen, and M. Poirié. Haemocyte changes in D. Melanogaster in response to long gland components of the parasitoid wasp Leptopilina boulardi: a Rho-GAP protein as an important factor. *Journal of insect physiology* 51.2 (Feb. 2005), 161–170.

[37] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9.4 (Mar. 2012), 357–359.

[38] S. Leclercq, J. Theze, M. A. Chebbi, I. Giraud, B. Moumen, L. Ernenwein, P. Grève, C. Gilbert, and R. Cordaux. Birth of a W sex chromosome by horizontal transfer of Wolbachia bacterial symbiont genome. *Proceedings of the National Academy of Sciences* 113.52 (Dec. 2016), 15036–15041.

[39] D. Lepetit, B. Gillet, S. Hughes, K. Kraaijeveld, and J. Varaldi. Genome Sequencing of the Behavior Manipulating Virus LbFV Reveals a Possible New Virus Family. *Genome biology and evolution* 8.12 (Dec. 2016), 3718–3739.

[40] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25.16 (Aug. 2009), 2078–2079.

[41] F. Louis, A. Bézier, G. Periquet, C. Ferras, J.-M. Drezen, and C. Dupuy. The bracovirus genome of the parasitoid wasp Cotesia congregata is amplified within 13 replication units, including sequences not packaged in the particles. *Journal of virology* 87.17 (Sept. 2013), 9649–9660.

[42] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)* 27.6 (Jan. 2011), 764–770.

[43] J. Martinez, F. Fleury, and J. Varaldi. Heritable variation in an extended phenotype: the case of a parasitoid manipulated by a virus. *Journal of Evolutionary Biology* 25.1 (Jan. 2012), 54–65.

[44] J. Martinez, F. Fleury, and J. Varaldi. Competitive outcome of multiple infections in a behavior-manipulating virus/wasp interaction. *Ecology and Evolution* 5.24 (Dec. 2015), 5934–45.

[45] P. Menzel, K. L. Ng, and A. Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* 7 (Apr. 2016), 1–9.

[46] B. Misof et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science (New York, NY)* 346.6210 (Nov. 2014), 763–767.

[47] J. Morales, H. Chiu, T. Oo, R. Plaza, S. Hoskins, and S. Govind. Biogenesis, structure, and immune-suppressive effects of virus-like particles of a Drosophila parasitoid, Leptopilina victoriae. *Journal of insect physiology* 51.2 (Feb. 2005), 181–195.

[48] N. Moran and T. Jarvik. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. *Science (New York, NY)* 328.5978 (Apr. 2010), 624–627.

[49] L. Mouton, H. Henri, M. Boulétreau, and F. Vavre. Strain-specific regulation of intracellular Wolbachia density in multiply infected insects. *Molecular Ecology* 12.12 (Dec. 2003), 3459–3465.

[50] N. Murphy, J. C. Banks, J. B. Whitfield, and A. D. Austin. Phylogeny of the parasitic microgastroid subfamilies (Hymenoptera: Braconidae) based on sequence data from seven genes, with an improved time estimate of the origin of the lineage. *Molecular phylogenetics and evolution* 47.1 (Apr. 2008), 378–395.

[51] M. M. van Oers and J. M. Vlak. Baculovirus genomics. *Current Drug Targets* 8.10 (Sept. 2007), 1051–1068.

[52] S. Patot, R. Allemand, F. Fleury, and J. Varaldi. An inherited virus influences the coexistence of parasitoid species through behaviour manipulation. *Ecology Letters* 15.6 (Apr. 2012), 603–610.

[53] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)* 28.11 (June 2012), 1420–1428.

[54] R. S. Peters et al. Evolutionary History of the Hymenoptera. *Current biology : CB* (Mar. 2017), 1–7.

[55] A. Pichon et al. Recurrent DNA virus domestication leading to different parasite virulence strategies. *Science Advances* 1.10 (Nov. 2015), e1501150–e1501150.

[56] M. Poirié, D. Colinet, and J.-L. Gatti. Insights into function and evolution of parasitoid wasp venoms. *Current Opinion in Insect Science* 6 (Dec. 2014), 52–60.

[57] R. M. Rizki and T. M. Rizki. Parasitoid virus-like particles destroy Drosophila cellular immunity. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 87.21 (Nov. 1990), 8388–8392.

[58] R. Saeedi, M. Li, and J. Frohlich. Clinical Biochemistry. *Clinical Biochemistry* 48.7-8 (May 2015), 472–475.

[59] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* 31.19 (Oct. 2015), 3210–3212.

[60] H. Sun, J. Ding, M. Piednoël, and K. Schneeberger. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics (Oxford, England)* 34.4 (Feb. 2018), 550–557.

[61] J. Theze, A. Bézier, G. Periquet, J.-M. Drezen, and E. A. Herniou. Paleozoic origin of insect large dsDNA viruses. *PNAS* 108.38 (Sept. 2011), 15931–15935.

[62] Y. H. Tzeng. Comparison of Three Methods for Estimating Rates of Synonymous and Nonsynonymous Nucleotide Substitutions. *Molecular biology and evolution* 21.12 (Aug. 2004), 2290–2298.

[63] J. Varaldi, P. Fouillet, M. Ravallec, M. Lopez-Ferber, M. Boulétreau, and F. Fleury. Infectious Behavior in a Parasitoid. *Science (New York, NY)* 302.5652 (Dec. 2003), 1930–1930.

[64] J. Varaldi and D. Lepetit. Deciphering the behaviour manipulation imposed by a virus on its parasitoid host: insights from a dual transcriptomic approach. *Parasitology* 3 (May 2018), 1–11.

[65] J. Varaldi, M. Ravallec, C. Labrosse, M. Lopez-Ferber, M. Boulétreau, and F. Fleury. Artifical transfer and morphological description of virus particles associated with superparasitism behaviour in a parasitoid wasp. *Journal of insect physiology* 52.11-12 (Oct. 2006), 1202–1212.

[66] A.-N. Volkoff, V. Jouan, S. Urbach, S. Samain, M. Bergoin, P. Wincker, E. Demettre, F. Cousserans, B. Provost, F. Coulibaly, F. Legeai, C. Béliveau, M. Cusson, G. Gyapay, and J.-M. Drezen. Analysis of virion structural components reveals vestiges of the ancestral ichnovirus genome. *PLoS pathogens* 6.5 (May 2010), e1000923.

[67] N. Wybouw, Y. Pauchet, D. G. Heckel, and T. Van Leeuwen. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome biology and evolution* 8.6 (June 2016), 1785–1801.

[68] H. Xu, X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* 7.12 (Dec. 2012), e52249.

[69] W. Xue, J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, and X.-W. Sun. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* 14.1 (Sept. 2013), 1–1.

7 Tables

basic statistics					BUSCO stats						Genome size [Mb]		
species	n_scaffolds	N50	coverage	Repetitive	Complete	Duplicated	Fragmented	Missing	total	missing	BUSCO.based	kmer.based	Cytometry.based
L. boulardi	127707	14511	46	27.65%	1044	4	8	10	1066	1%	353	347	361
L. heterotoma	231242	12807	53	28.82 %	1041	2	9	14	1066	1%	445	464	459
L. clavipes	38495	17657	83	24.02 %	1025	7	15	19	1066	2%	257	300	321
G. brasiliensis	2777766	2080	24	44.92 %	830	8	192	36	1066	3%	829	977	968

Table 1: Statistics for the assemblies of wasp genomes. Genome size was estimated either using the coverage on BUSCO gene containing scaffolds or using a k-mer approach. For comparison, we give the estimated genome sizes obtained from flow cytometry analysis [23][35].

query			*L. boulardi*			*L. heterotoma*			*L. clavipes*		
	query_id	Length	identity	aln.length	evalue	identity	aln.length	evalue	identity	aln.length	evalue
1	LbFV_ORF5	696	34.40	366	5.5e-41	29.70	370	3e-37	33.10	366	1.9e-40
2	LbFV_ORF72	106	31.80	107	5.2e-10	28.60	70	4e-04	32.70	107	8.8e-09
3	LbFV_ORF92	1593	33.80	1058	2.9e-151	38.10	501	5e-94	33.70	998	3.1e-136
4	LbFV_ORF107	625	29.80	322	1.3e-11	27.10	170	9e-09	28.30	378	5.3e-10
5	LbFV_ORF94	182	29.00	176	5.5e-14	27.60	174	1e-11	27.00	174	1.2e-12
6	LbFV_ORF68	645	34.10	646	6.7e-99	32.60	660	3e-92	34.00	674	3.5e-103
7	LbFV_ORF60	362	32.60	377	2.4e-36	26.00	381	7e-30	31.80	384	1.4e-33
8	LbFV_ORF85	215	36.40	225	3.0e-26	35.20	219	1e-23	33.00	218	1.3e-23
9	LbFV_ORF87	176	30.90	162	6.5e-12	29.00	162	1e-05	31.50	165	3.6e-11
10	LbFV_ORF58	1308	36.70	932	1.3e-129	31.50	1378	8e-158	31.50	1042	1.8e-120
11	LbFV_ORF78	676	40.10	670	1.2e-134	41.00	646	2e-123	41.00	675	3.7e-135
12	LbFV_ORF83	433	24.80	435	1.6e-15	21.90	429	8e-15	24.50	436	1.8e-20
13	LbFV_ORF96	1048	41.90	1024	4.0e-169	36.60	1043	2e-164	40.40	1013	1.3e-178

Table 2: Blast hits for the 13 viral proteins against *Leptopilina* genomes (tblastn).

locus	species	alignment.start	alignment.end	envelope.start	envelope.end	accession	family name	hmm.start	hmm.end	hmm.length	bit.score	Individual.E.value	Conditional.E.value
ORF58	LbFV	639	870	599	880	PF00136.20	DNA_pol_B	40	200	464	72.63	2.3e-20	1.4e-24
ORF58	L. clavipes	349	578	322	591	PF00136.20	DNA_pol_B	19	205	464	23.88	1.4e-05	1.7e-09
ORF60	LbFV	76	172	57	351	PF02450.14	LCAT	66	165	392	30.75	1.6e-07	6.7e-11
ORF60	L. boulardi	121	218	105	234	PF02450.14	LCAT	76	172	392	25.45	6.6e-06	3.5e-09
ORF60	L. heterotoma	120	218	103	284	PF02450.14	LCAT	76	173	392	27.26	1.8e-06	9.9e-10
ORF60	L. clavipes	120	367	103	398	PF02450.14	LCAT	76	280	392	25.24	7.6e-06	4.1e-09
ORF68	LbFV	124	167	122	174	PF05970.13	PIF1-like helicase	3	46	364	21.87	8.0e-05	3.3e-08
ORF68	LbFV	248	320	226	379	PF05970.13	PIF1-like helicase	103	171	364	15.24	8.3e-03	3.5e-06
ORF68	L. boulardi	138	181	138	191	PF05970.13	PIF1-like helicase	1	44	364	11.92	8.4e-02	7.6e-05
ORF68	L. boulardi	273	344	261	388	PF05970.13	PIF1-like helicase	104	175	364	11.54	1.1e-01	9.8e-05
ORF68	L. heterotoma	139	182	139	193	PF05970.13	PIF1-like helicase	1	44	364	11.49	1.1e-01	8.9e-05
ORF68	L. heterotoma	283	353	260	396	PF05970.13	PIF1-like helicase	104	174	364	16.27	4.0e-03	3.1e-06
ORF68	L. clavipes	142	183	141	193	PF05970.13	PIF1-like helicase	2	43	364	8.51	9.2e-01	8.8e-04
ORF68	L. clavipes	284	339	265	358	PF05970.13	PIF1-like helicase	103	158	364	12.71	4.8e-02	4.6e-05
ORF78	LbFV	358	415	244	422	PF00623.19	RNA_pol_Rpb1_2	100	156	166	16.14	9.1e-03	5.4e-07
ORF78	L. boulardi	238	299	232	303	PF00623.19	RNA_pol_Rpb1_2	100	160	166	15.16	1.8e-02	1.1e-06
ORF78	L. heterotoma	206	273	149	277	PF00623.19	RNA_pol_Rpb1_2	95	161	166	18.21	2.1e-03	1.2e-07
ORF78	L. clavipes	236	305	202	309	PF00623.19	RNA_pol_Rpb1_2	93	161	166	19.14	1.1e-03	1.3e-07
ORF85	LbFV	56	201	5	201	PF05820.10	Ac81	28	181	181	77.15	1.1e-21	1.3e-25
ORF85	L. boulardi	62	214	41	214	PF05820.10	Ac81	26	181	181	74.16	9.0e-21	1.1e-24
ORF85	L. heterotoma	63	213	34	213	PF05820.10	Ac81	29	181	181	78.91	3.1e-22	3.7e-26
ORF85	L. clavipes	59	212	34	212	PF05820.10	Ac81	25	181	181	73.61	1.3e-20	7.9e-25

Table 3: hmmer sequence analysis for the 13 proteins encoded by LbFV and their orthologs in *Leptopilina* wasps. Only hits with individual evalues < 0.15 are shown.

8 Figures

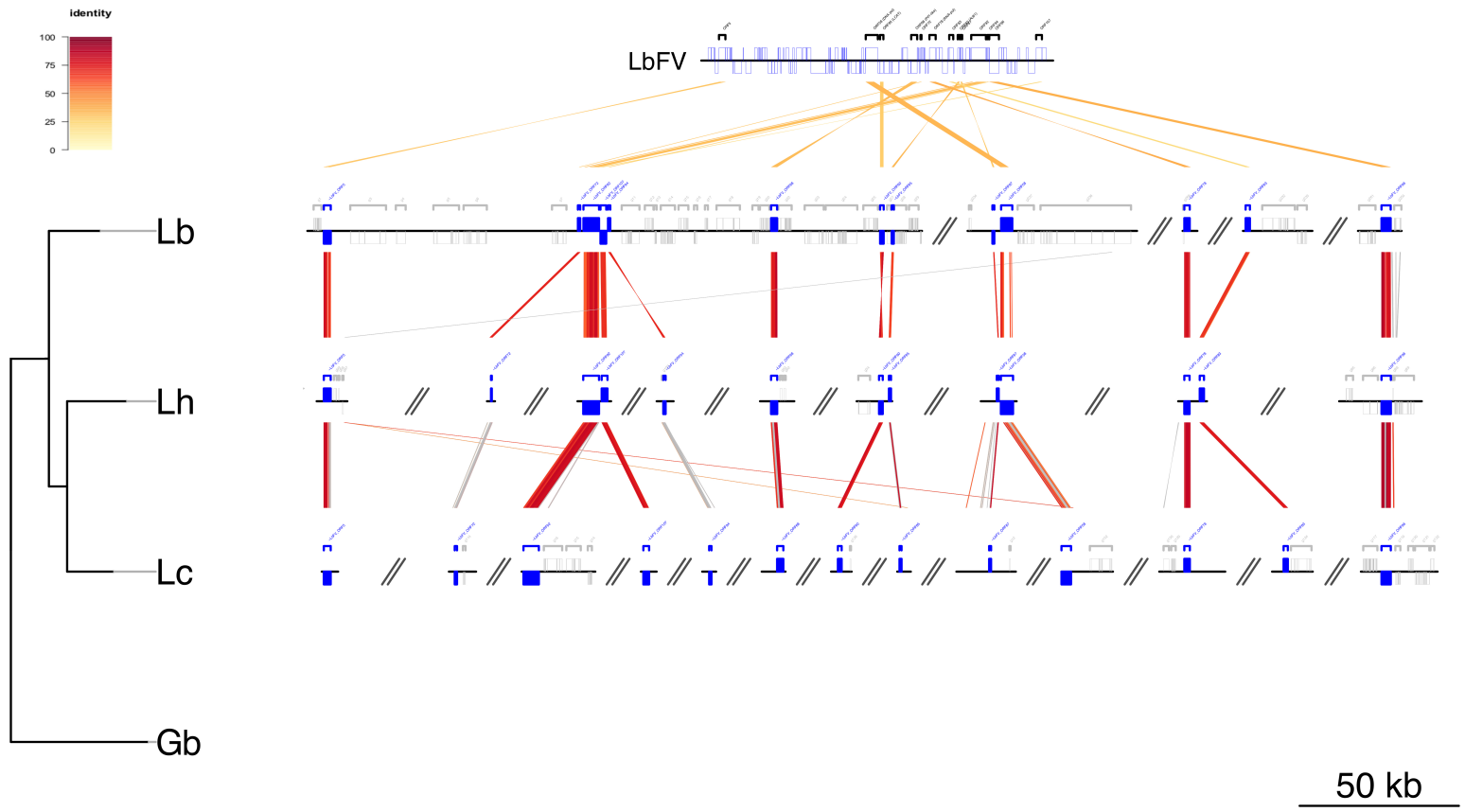

Figure 1: Comparative genomics of wasp scaffolds sharing similarities with virus proteins. Lb: *L. boulardi*, Lh: *L. heterotoma*, Lc: *L. clavipes*, Gb: *Ganaspis brasiliensis*, LbFV: Leptopilina boulardi Filamentous Virus. (A) The species-tree on the left has been obtained using a concatenation of 627 universal arthropod genes. All branches (Lh-Lc and Lh-Lc/Lb) have an aLRT value of 1 (*Apis mellifera* was used as an outgroup). The red/yellow color code depicts the percentage of protein identity between homologous sequence pairs (viral or virally-derived loci). Grey connections indicate homology between regions that does not contain virally-derived loci. Genes of eukaryotic origin are depicted in grey on the scaffolds. The figure has been drawn using the genoPlotR package[28]. The scaffolds are ordered from left to right in an arbitrary manner.

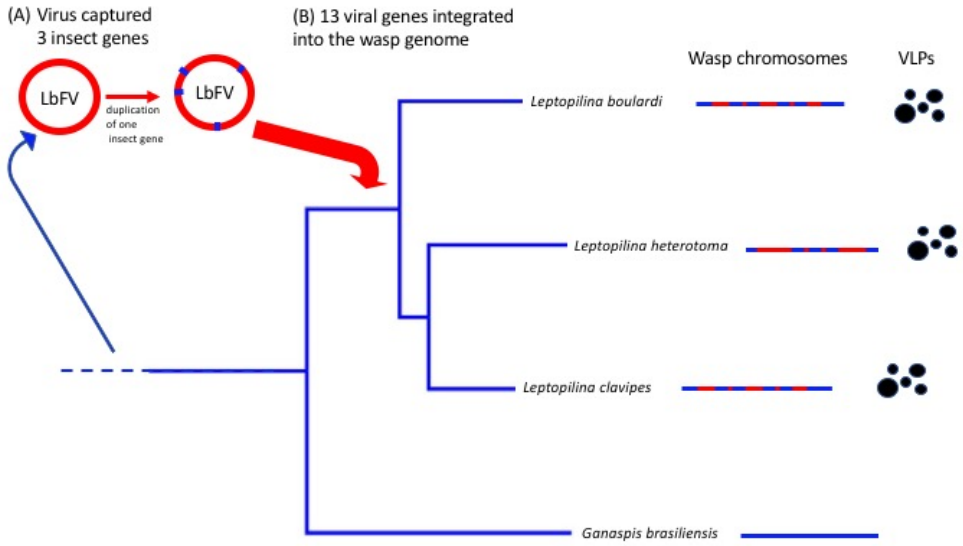

Figure 2: Hypothetical scenario for genetic exchanges between the wasps and the virus LbFV. (A) Before the diversification of the *Leptopilina* genus, LbFV captured 3 insect genes, most likely involved in apoptosis inhibition (ORFs 27 and 66) and methylation (the ancestor of ORFs 11 and 13). One of them was probably subsequently duplicated (the ancestor of both ORFs 11 and 13). (B) After the divergence between *Ganaspis* and *Leptopilina* (around 74My ago[8]), but before the diversification of *Leptopilina* genus, possibly a whole genome of a virus closely related to LbFV integrated wasp chromosomes. Nowadays, all *Leptopilina* species bear 13 LbFV-derived genes that allow them to produce VLPs. The cartoons displaying the chromomosomes are just illustrations depicting the presence of virally-derived genes (red) within wasp chromosomes of eukaryotic origin (blue). VLPs are symbolized by the black circular forms.

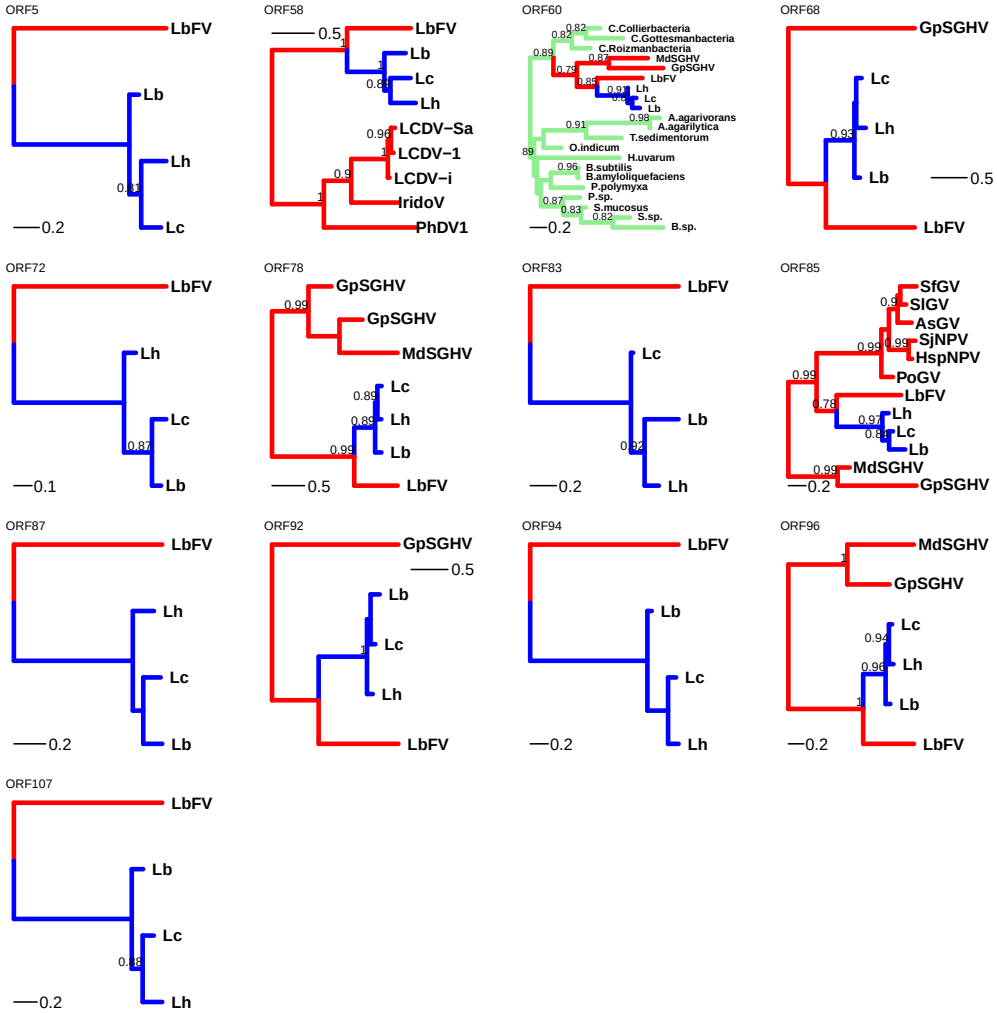

Figure 3: Phylogenetic evidence for a massive horizontal transfer of thirteen viral genes into the genome of *Leptopilina* wasps. The names of the ORFs refers to the ORF number in LbFV genome. Blue, red and green colors represent respectively (supposedly) eukaryotic, viral or bacterial branches. Only aLRT supports > 0.7 are shown. The mid-point rooting method was used. Accession numbers of the corresponding sequences are available in table S4.

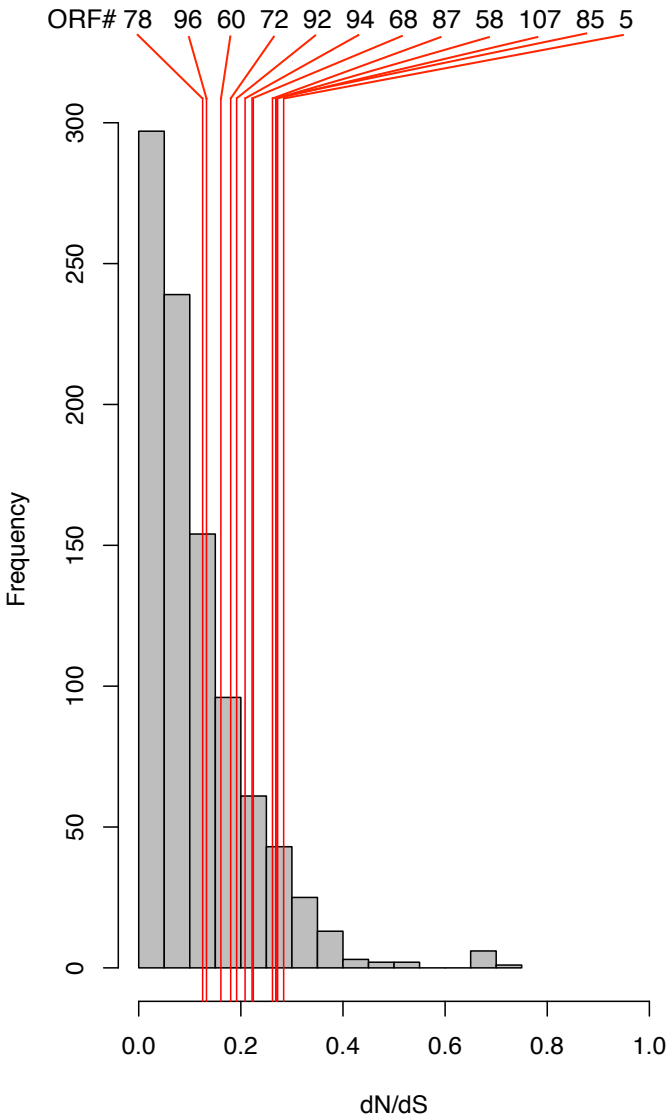

Figure 4: dN/dS ratio for a set of 942 universal arthropod genes and for the 13 virally derived genes found in *Leptopilina* species (indicated by the red lines).

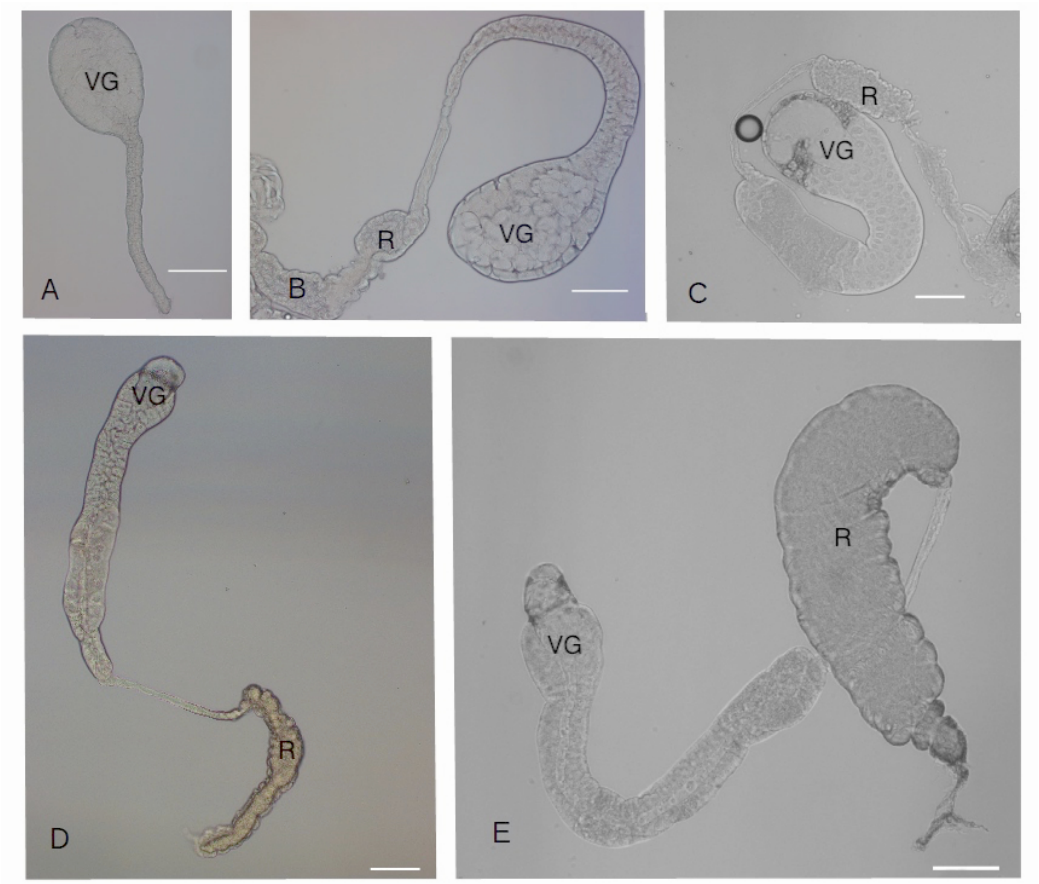

Figure 5: Morphogenesis of the venom gland during the pupal stage of *L. boulardi* females. G: venom gland; R: reservoir of the venom gland. Overall structure of the organ under light microscope at day 11 (a), 14 (b), 16 (c), 18(d) and 21(e). At that temperature (25°C), 11 days corresponds to the beginning of the pupation in *L. boulardi*, whereas adult females are emerging at 21 days. Bar= 100μM.

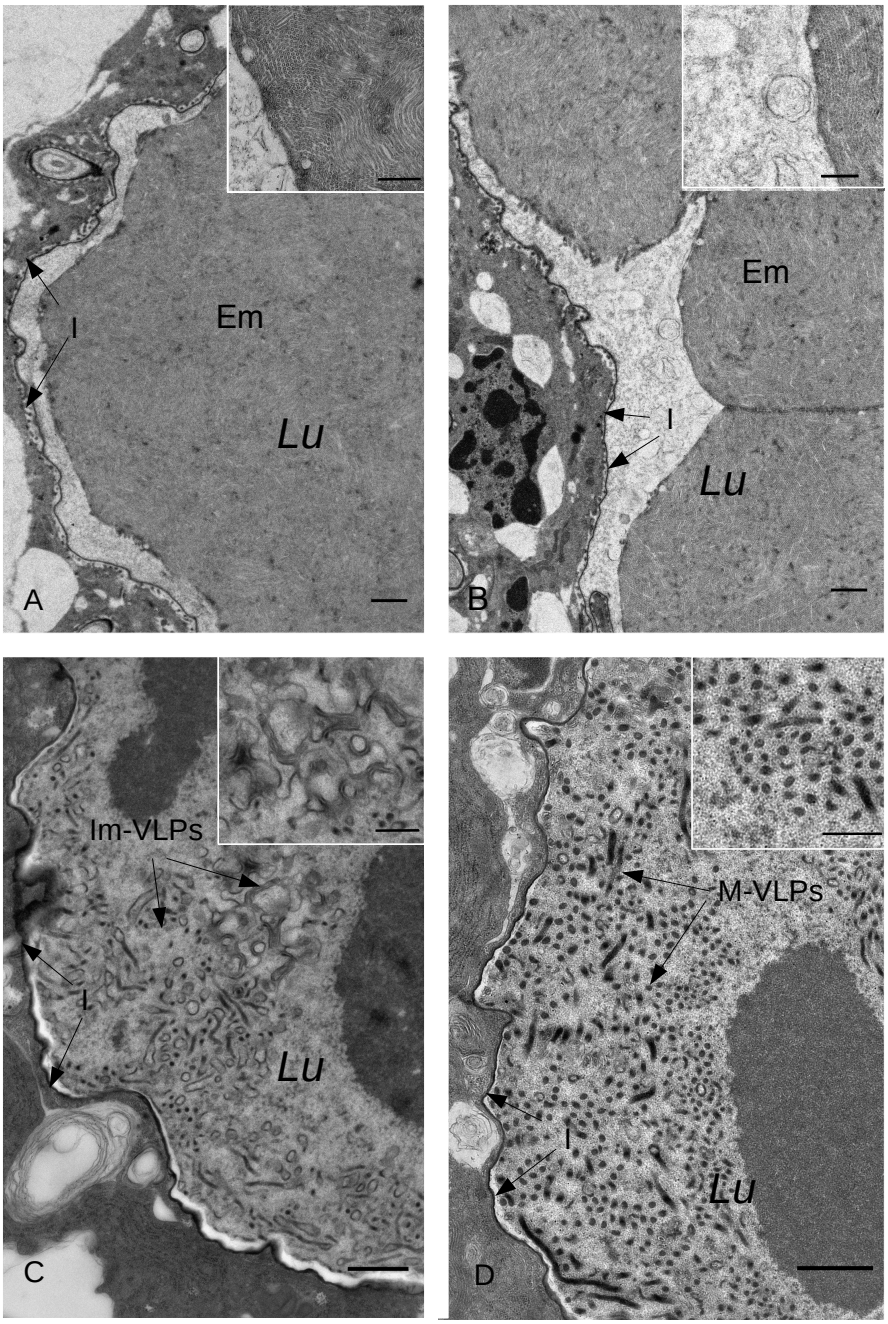

Figure 6: Biogenesis of VLPs in the venom gland of *L. boulardi* during the pupal stage until adult emergence: (A) 14 days (pupae), (B) 16 days (pupae), (C) 18 days (pupae), (D) 21 days (adult). At days 14 and 16, secretory cells (SC) are releasing empty membranes (Em) into the Lumen (Lu) of the venom gland where they accumulate. Then at day 18, empty membranes starts to be filled with electron-dense material (probably virulence proteins, such as LbGAP) to produce immature VLPs (im-VLPs). Finally at emergence (day 21), the venom gland lumen is filled with mature VLPs (m-VLPs) ready to be injected into the host. I: cuticular intima delineating the lumen. Inserts show details of each image. Bars represent 1μM, except in inserts where they represent 500μM.

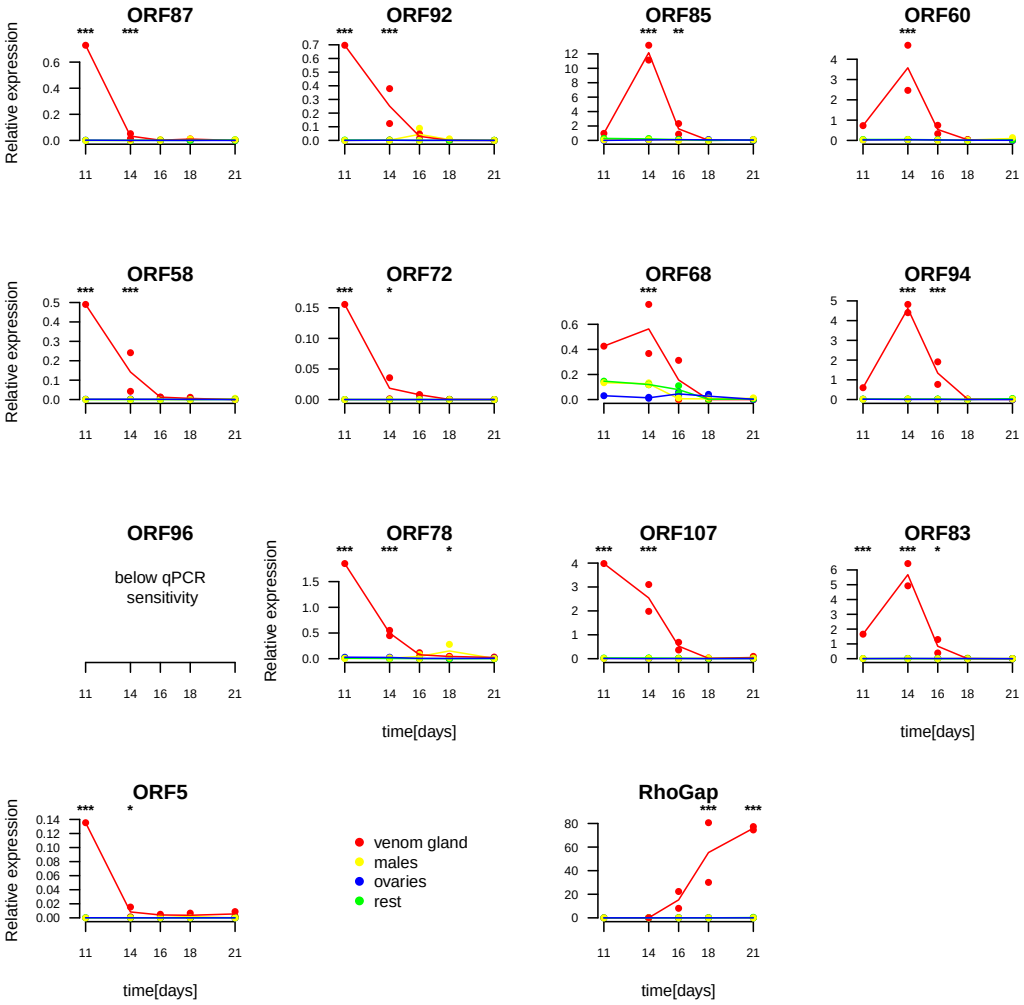

Figure 7: Expression of the 13 virally-derived genes and of the Rho-Gap in different tissues of *L. boulardi* from initial pupal stage to adult. x-axis represents days since egg-laying. 11 days corresponds to the beginning of the pupal stage and 21 days to the emergence of adults from the *Drosophila* puparium.

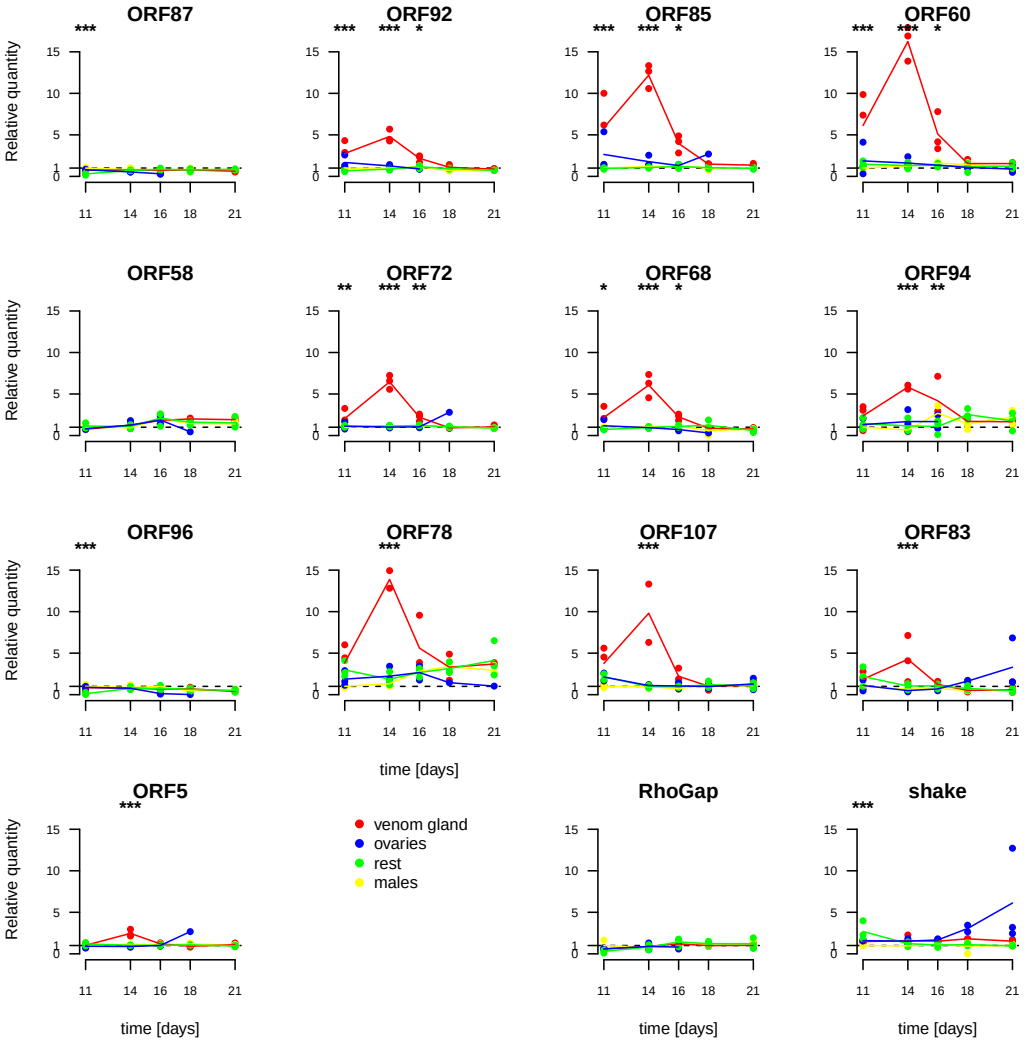

Figure 8: Genomic amplification of virally derived-genes measured by real time PCR in *L. boulardi*. The relative quantity of each target gene is represented relative to the actin control gene and normalized by the ratio observed in males at day 11. The expected value under no amplification (relative quantity=1) is indicated as a dotted line. Stars correspond to the tissue effect tested at each time point (with holm correction for multiple tests) : $* < 0.05$, $** < 0.01$, $*** < 0.001$.

9 Supplementary tables and figures

	blast output											corresponding ORF on scaffold				scaffold statistics		
	query_id	query_len	subject_id	identity	aln length	qstart	qend	sstart	send	evalue	bitscore	start	end	length	strand	scaf_length	cov_depth	GC
1	LbFV_ORF5	696	scaffold_159	34.4	366	337	696	6401	5337	5.5e-41	164.00	7601	5337	755	+	435056	53	0.36
2	LbFV_ORF72	106	scaffold_159	31.8	107	2	102	88433	88753	5.2e-10	57.40	88025	88771	249	-	435056	53	0.36
3	LbFV_ORF92	1593	scaffold_159	33.8	1058	583	1593	91842	94901	2.9e-151	518.00	89832	94901	1690	-	435056	53	0.36
4	LbFV_ORF107	625	scaffold_159	29.8	322	320	625	96312	95377	1.3e-11	71.20	97248	95377	624	+	435056	53	0.36
5	LbFV_ORF94	182	scaffold_159	29.0	176	1	173	98066	98557	5.5e-14	72.00	97829	98569	247	-	435056	53	0.36
6	LbFV_ORF68	645	scaffold_159	34.1	646	29	642	150985	152847	6.7e-99	335.00	150889	152856	656	-	435056	53	0.36
7	LbFV_ORF60	362	scaffold_159	32.6	377	5	353	187445	186375	2.4e-36	143.00	187532	186366	389	+	435056	53	0.36
8	LbFV_ORF85	215	scaffold_159	36.4	225	1	212	190829	190170	3.0e-26	108.00	190829	190149	227	+	435056	53	0.36
9	LbFV_ORF87	176	scaffold_2503	30.9	162	8	158	8659	8183	6.5e-12	65.90	8698	8078	207	+	55139	44	0.22
10	LbFV_ORF58	1308	scaffold_2503	36.7	932	3	904	10711	13299	1.3e-129	446.00	10909	14550	1214	-	55139	44	0.22
11	LbFV_ORF78	676	IDBA_scaffold_13958	40.1	670	43	670	2268	4205	1.2e-134	434.00	2487	4241	585	-	4800	49	0.57
12	LbFV_ORF83	433	scaffold_2315	24.8	435	14	407	874	2139	1.6e-15	82.40	862	2259	466	-	22591	45	0.20
13	LbFV_ORF96	1048	IDBA_scaffold_2184	41.9	1024	48	1041	3609	6512	4.0e-169	554.00	3564	6545	994	-	14197	45	0.28

Table S1: Blast hits for the 13 viral genes against *L. boulardi* genome.

	blast output											corresponding ORF on scaffold				scaffold statistics		
	query_id	query_len	subject_id	identity	aln length	qstart	qend	sstart	send	evalue	bitscore	start	end	length	strand	scaf_length	cov_depth	GC
1	LbFV_ORF5	696	IDBA_scaffold_8257	29.7	370	333	696	6582	7661	3e-37	157.00	5424	7661	746	-	9987	59	0.29
2	LbFV_ORF72	106	IDBA_scaffold_32827	28.6	70	34	102	1541	1750	4e-04	36.60	1303	1563	87	-	2607	58	0.23
3	LbFV_ORF92	1593	IDBA_scaffold_7081	38.1	501	1109	1590	5437	3938	5e-94	347.00	9070	3929	1714	+	10934	53	0.29
4	LbFV_ORF107	625	IDBA_scaffold_7081	27.1	170	455	621	2550	3056	9e-09	62.40	1179	3065	629	-	10934	53	0.29
5	LbFV_ORF94	182	IDBA_scaffold_13988	27.6	174	1	171	2671	2186	1e-11	69.70	2905	2168	246	+	5494	53	0.23
6	LbFV_ORF68	645	IDBA_scaffold_6001	32.6	660	29	644	7459	5555	3e-92	339.00	7561	5552	670	+	11133	52	0.48
7	LbFV_ORF60	362	scaffold_1324	26.0	381	5	353	4186	3065	7e-30	131.00	4270	3056	405	+	11549	50	0.34
8	LbFV_ORF85	215	scaffold_1324	35.2	219	1	207	375	1031	1e-23	109.00	375	1052	226	-	11549	50	0.34
9	LbFV_ORF87	176	IDBA_scaffold_5653	29.0	162	8	162	5879	6355	1e-05	49.70	5834	6457	208	-	11655	53	0.32
10	LbFV_ORF58	1308	IDBA_scaffold_5653	31.5	1378	19	1299	5204	1260	8e-158	558.00	5126	1170	1319	+	11655	53	0.32
11	LbFV_ORF78	676	IDBA_scaffold_9791	41.0	646	70	669	3914	2034	2e-123	443.00	3692	1992	567	+	9362	52	0.21
12	LbFV_ORF83	433	IDBA_scaffold_9791	21.9	429	14	407	7018	8277	8e-15	82.00	7006	8385	460	-	9362	52	0.21
13	LbFV_ORF96	1048	IDBA_scaffold_1712	36.6	1043	48	1041	16775	13806	2e-164	580.00	16820	13773	1016	+	26871	53	0.29

Table S2: Blast hits for the 13 viral genes against *L. heterotoma* genome.

	blast output											corresponding ORF on scaffold				scaffold statistics		
	query_id	query_len	subject_id	identity	aln length	qstart	qend	sstart	send	evalue	bitscore	start	end	length	strand	scaf_length	cov_depth	GC
1	LbFV_ORF5	696	scf7180005159507	33.1	366	337	696	1730	663	1.9e-40	162.00	2906	663	748	+	5318	87	0.31
2	LbFV_ORF72	106	scf7180005166731	32.7	107	2	102	6537	6217	8.8e-09	53.90	6945	6199	249	+	8832	81	0.30
3	LbFV_ORF92	1593	scaffold_1017	33.7	998	579	1536	21309	18403	3.1e-136	472.00	23376	18370	1669	+	23961	75	0.27
4	LbFV_ORF107	625	scf7180005156365	28.3	378	265	622	1897	809	5.3e-10	65.50	2674	803	624	+	5122	96	0.30
5	LbFV_ORF94	182	scf7180005161552	27.0	174	1	171	2763	2278	1.2e-12	67.80	3015	2260	252	+	4524	62	0.28
6	LbFV_ORF68	645	scf7180005174277	34.0	674	29	644	5118	7034	3.5e-103	347.00	5016	7037	674	-	7741	213	0.30
7	LbFV_ORF60	362	scf7180005174113	31.8	384	5	353	2297	3421	1.4e-33	134.00	2213	3430	406	-	6683	57	0.29
8	LbFV_ORF85	215	scf7180005171671	33.0	218	1	207	3017	3670	1.3e-23	100.00	3017	3691	225	-	4425	83	0.29
9	LbFV_ORF87	176	scaffold_886	31.5	165	8	158	8088	8570	3.6e-11	63.20	8049	8678	210	-	19330	85	0.28
10	LbFV_ORF58	1308	scf7180005154334	31.5	1042	317	1288	16626	13723	1.8e-120	418.00	16746	13633	1038	+	16768	70	0.26
11	LbFV_ORF78	676	scf7180005177077	41.0	675	39	669	11274	13268	3.7e-135	441.00	11517	13316	600	-	21465	86	0.29
12	LbFV_ORF83	433	scf7180005174071	24.5	436	9	404	3734	5005	1.8e-20	97.40	3740	5122	461	-	13231	85	0.28
13	LbFV_ORF96	1048	scf7180005173345	40.4	1013	48	1021	9667	6782	1.3e-178	582.00	9712	6686	1009	+	24926	74	0.28

Table S3: Blast hits for the 13 viral genes against *L. clavipes* genome.

	Locus	species	GI	Figure
1	ORF5	Lb	PQAT00000000	3
2	ORF5	Lh	pending acc. num.	3
3	ORF5	Lc	JUFY01000000	3
4	ORF5	LbFV	1148998810	3
5	ORF58	Lb	PQAT00000000	3
6	ORF58	Lh	pending acc. num.	3
7	ORF58	Lc	JUFY01000000	3
8	ORF58	LbFV	1148998708	3
9	ORF60	Lb	PQAT00000000	3
10	ORF60	Lh	pending acc. num.	3
11	ORF60	Lc	JUFY01000000	3
12	ORF60	Lymphocystis_disease_virus_-_isolate_China	51870153	3
13	ORF60	Organic_Lake_phycodnavirus_1	322510829	3
14	ORF60	Invertebrate_iridovirus_25	589287870	3
15	ORF60	Lymphocystis_disease_virus_1	611962711	3
16	ORF60	Lymphocystis_disease_virus_Sa	1135106808	3
17	ORF60	LbFV	1148998761	3
18	ORF68	Acyrthosiphon_pisum	328698707	3
19	ORF68	Adoxophyes_honmai_entomopoxvirus_L	506498063	3
20	ORF68	Apis_cerana_cerana	1241837182	3
21	ORF68	Apis_dorsata	572314547	3
22	ORF68	Apis_florea	820863019	3
23	ORF68	Apis_mellifera	571506210	3
24	ORF68	Bombus_terrestris	340708910	3
25	ORF68	Camponotus_floridanus	752871224	3
26	ORF68	Cephus_cinctus	1000753753	3
27	ORF68	Chlamydotis_macqueenii	677160893	3
28	ORF68	Crassostrea_gigas	1139814932	3
29	ORF68	Cuculus_canorus	676590237	3
30	ORF68	Dendroctonus_ponderosae	546685733	3
31	ORF68	Diaphorina_citri	662192917	3
32	ORF68	Diuraphis_noxia	985403395	3
33	ORF68	Dufourea_novaeangliae	987914045	3
34	ORF68	Eufriesea_mexicana	1059214553	3

35	ORF68	Glossina_morsitans_morsitans	83595237	3
36	ORF68	Gb	pending acc. num.	3
37	ORF68	Habropoda_laboriosa	1059864473	3
38	ORF68	Harpegnathos_saltator	749795708	3
39	ORF68	Helicoverpa_armigera	304423112	3
40	ORF68	Lasius_niger	861651735	3
41	ORF68	Lb	_PQAT00000000	3
42	ORF68	LbFV	1148998769	3
43	ORF68	Lc	JUFY01000000	3
44	ORF68	Lh	pending acc. num.	3
45	ORF68	Myzus_persicae	1230193237	3
46	ORF68	Nasonia_vitripennis	1032757220	3
47	ORF68	Opisthocomus_hoazin	677549512	3
48	ORF68	Papilio_machaon	930680047	3
49	ORF68	Papilio_xuthus	910339325	3
50	ORF68	Parasteatoda_tepidariorum	1009572498	3
51	ORF68	Pogonomyrmex_barbatus	769838565	3
52	ORF68	Polistes_canadensis	954577453	3
53	ORF68	Trichomalopsis_sarcophagae	1227108847	3
54	ORF68	Trichoplusia_ni	6635437	3
55	ORF68	Vollenhovia_emeryi	795079157	3
56	ORF72	Lb_LbFV	pending acc. num.	3
57	ORF72	Lh_LbFV	pending acc. num.	3
58	ORF72	Lc_LbFV	pending acc. num.	3
59	ORF72	Glossina_pallidipes_salivary_gland_hypertrophy_virus	168804090	3
60	ORF72	LbFV	1148998771	3
61	ORF78	Lb	PQAT00000000	3
62	ORF78	Lh	pending acc. num.	3
63	ORF78	Lc	JUFY01000000	3
64	ORF78	LbFV	1148998775	3
65	ORF83	Lb_LbFV	pending acc. num.	3
66	ORF83	Lh_LbFV	pending acc. num.	3
67	ORF83	Lc_LbFV	pending acc. num.	3
68	ORF83	Musca_domestica_salivary_gland_hypertrophy_virus	187903111	3
69	ORF83	Glossina_pallidipes_salivary_gland_hypertrophy_virus	984290647	3
70	ORF83	Glossina_pallidipes_salivary_gland_hypertrophy_virus	984290648	3
71	ORF83	LbFV	1148998781	3
72	ORF85	Lb	PQAT00000000	3
73	ORF85	Lh	pending acc. num.	3
74	ORF85	Lc	JUFY01000000	3
75	ORF85	LbFV	1148998786	3
76	ORF87	Lb	PQAT00000000	3
77	ORF87	Lh	pending acc. num.	3
78	ORF87	Lc	JUFY01000000	3
79	ORF87	Phthorimaea_operculella_granulovirus	21686761	3
80	ORF87	Agrotis_segetum_granulovirus	46309360	3
81	ORF87	Spodoptera_litura_granulovirus	148368915	3
82	ORF87	Glossina_pallidipes_salivary_gland_hypertrophy_virus	168804094	3
83	ORF87	Musca_domestica_salivary_gland_hypertrophy_virus	187903145	3
84	ORF87	Hemileuca_sp._nucleopolyhedrovirus	529218126	3
85	ORF87	Spodoptera_frugiperda_granulovirus	761719624	3
86	ORF87	Sucra_jujuba_nucleopolyhedrovirus	960494866	3
87	ORF87	Glossina_pallidipes_salivary_gland_hypertrophy_virus	984290700	3
88	ORF87	LbFV	1148998788	3
89	ORF92	Lb	PQAT00000000	3
90	ORF92	Lh	pending acc. num.	3
91	ORF92	Lc	JUFY01000000	3
92	ORF92	LbFV	1148998790	3
93	ORF94	Lb_LbFV	pending acc. num.	3
94	ORF94	Lh_LbFV	pending acc. num.	3
95	ORF94	Lc_LbFV	pending acc. num.	3
96	ORF94	Glossina_pallidipes_salivary_gland_hypertrophy_virus	168804177	3
97	ORF94	LbFV	1148998795	3
98	ORF96	Lb	PQAT00000000	3
99	ORF96	Lh	pending acc. num.	3
100	ORF96	Lc	JUFY01000000	3
101	ORF96	LbFV	1148998797	3
102	ORF107	Lb_LbFV	pending acc. num.	3
103	ORF107	Lh_LbFV	pending acc. num.	3
104	ORF107	Lc_LbFV	pending acc. num.	3
105	ORF107	Glossina_pallidipes_salivary_gland_hypertrophy_virus	168804057	3
106	ORF107	Musca_domestica_salivary_gland_hypertrophy_virus	187903107	3
107	ORF107	LbFV	1148998799	3
108	ITS2	L.longipes	AF015893.1	S4
109	ITS2	L.guineaensis	AY124559.1	S4
110	ITS2	L.victoriae	AY124553.1	S4
111	ITS2	L.heterotoma	AB546896.1	S4
112	ITS2	L.orientalis	AY124563.1	S4
113	ITS2	L.boulardi	AY124568.1	S4
114	ITS2	L.freyae	AY124561.1	S4
115	ITS2	L.fimbriata	AF015894.1	S4
116	ITS2	L.clavipes	JQ808416.1	S4
117	ITS2	L.australis	AF015897.1	S4
118	ITS2	G.brasiliensis	AB678777.1	S4
1	ORF27	Papilio xuthus	XP_013173302.1	S1A
2	ORF27	Bicyclus anynana	XP_023937808.1	S1A
3	ORF27	Pieris rapae	XP_022114989.1	S1A
4	ORF27	Spodoptera litura	XP_022828254.1	S1A
5	ORF27	Bombyx mori	NP_001037024.1	S1A
6	ORF27	Drosophila busckii	XP_017843635.1	S1A
7	ORF27	Musca domestica	XP_005178734.1	S1A
8	ORF27	Zeugodacus cucurbitae	XP_011180685.1	S1A
9	ORF27	Ceratitis capitata	XP_004519914.1	S1A
10	ORF27	Dendroctonus ponderosae	XP_019755885.1	S1A
11	ORF27	Anoplophora glabripennis	XP_018566786.1	S1A
12	ORF27	Leptinotarsa decemlineata	XP_023022306.1	S1A
13	ORF27	Polistes dominula	XP_015178412.1	S1A
14	ORF27	Linepithema humile	XP_012229104.1	S1A
15	ORF27	Camponotus floridanus	XP_011252805.1	S1A
16	ORF27	Pogonomyrmex barbatus	XP_011630441.1	S1A
17	ORF27	Megachile rotundata	XP_012151451.1	S1A
18	ORF27	Microplitis demolitor	XP_008554575.1	S1A
19	ORF27	Fopius arisanus	XP_011298329.1	S1A
20	ORF27	Diachasma alloeum	XP_015109162.1	S1A
21	ORF27	Cephus cinctus	XP_015599785.1	S1A
22	ORF27	Ganaspis brasiliensis	pending_NA	S1A
23	ORF27	Leptopilina boulardi	PQAT00000000.1_NA	S1A
24	ORF27	Leptopilina heterotoma	pending_NA	S1A
25	ORF27	Leptopilina clavipes	JUFY00000000.1_NA	S1A
26	ORF27	Orussus abietinus	XP_012276925.1	S1A

27	ORF27	Nasonia vitripennis	XP_016838993.1	S1A
28	ORF27	LbFV	1148998730	S1A
29	ORF27	Dufourea novaeangliae	XP_015432901.1	S1A
30	ORF27	Apis florea	XP_012348205.1	S1A
31	ORF27	Apis mellifera	XP_006570777.1	S1A
32	ORF27	Habropoda laboriosa	XP_017799036.1	S1A
33	ORF27	Bombus terrestris	XP_012163415.1	S1A
34	ORF66	Harpegnathos saltator	749795708	S1B
35	ORF66	Camponotus floridanus	752871224	S1B
36	ORF66	Pogonomyrmex barbatus	769838565	S1B
37	ORF66	Vollenhovia emeryi	795079157	S1B
38	ORF66	Nasonia vitripennis	1032757220	S1B
39	ORF66	Trichomalopsis sarcophagae	1227108847	S1B
40	ORF66	Cephus cinctus	1000753753	S1B
41	ORF66	Ganaspis brasiliensis	pending	S1B
42	ORF66	Leptopilina heterotoma	pending	S1B
43	ORF66	Leptopilina clavipes	JUFY00000000.1	S1B
44	ORF66	Leptopilina boulardi	PQAT00000000.1	S1B
45	ORF66	Dufourea novaeangliae	987914045	S1B
46	ORF66	Habropoda laboriosa	1059864473	S1B
47	ORF66	Apis florea	820863019	S1B
48	ORF66	Apis dorsata	572314547	S1B
49	ORF66	Apis mellifera	571506210	S1B
50	ORF66	Apis cerana cerana	1241837182	S1B
51	ORF66	Eufriesea mexicana	1059214553	S1B
52	ORF66	Bombus terrestris	340708910	S1B
53	ORF66	Agrilus planipennis	XP_018331076.1	S1B
54	ORF66	Tribolium castaneum	NP_001280519.1	S1B
55	ORF66	Nicrophorus vespilloides	XP_017784576.1	S1B
56	ORF66	Papilio machaon	930680047	S1B
57	ORF66	Papilio xuthus	910339325	S1B
58	ORF66	Helicoverpa armigera	304423112	S1B
59	ORF66	Trichoplusia ni	6635437	S1B
60	ORF66	Acyrthosiphon pisum	328698707	S1B
61	ORF66	Diuraphis noxia	985403395	S1B
62	ORF66	Myzus persicae	1230193237	S1B
63	ORF66	LbFV	1148998769	S1B
64	ORF66	Adoxophyes honmai EPV	506498063	S1B
65	ORF11-13	LbFVorf11	009345615	S1C
66	ORF11-13	Ganaspis brasiliensis	pending	S1C
67	ORF11-13	LbFVorf13	009345617.1	S1C
68	ORF11-13	Leptopilina boulardi	PQAT00000000.1	S1C
69	ORF11-13	Leptopilina heterotoma	pending	S1C
70	ORF11-13	Leptopilina clavipes	JUFY00000000.1	S1C
71	ORF11-13	Exserohilum turcica	XP_008030043.1	S1C
72	ORF11-13	Alternaria alternata	XP_018379425.1	S1C
73	ORF11-13	Frankliniella occidentalis	XP_026288761.1	S1C
74	ORF11-13	Rhizopus microsporus	XP_023462188.1	S1C
75	ORF11-13	Melampsora larici-populina	XP_007414376.1	S1C
76	ORF11-13	Debaryomyces hansenii	XP_459998.2	S1C
77	ORF11-13	Debaryomyces fabryi	XP_015465751.1	S1C
78	ORF11-13	Eremothecium gossypii	NP_986783.2	S1C
79	ORF11-13	Eremothecium cymbalariae	XP_003645815.1	S1C

Table S4: Accession numbers of sequences used in the phylogenies

	primer_name	Orientation	tm	GC	Seq	Prod.Size
1	Lb_ORF96_F	FORWARD	59.99	55	AATGGAGGACTACCGACACG	259
2	Lb_ORF96_R	REVERSE	59.62	47	TGCACTGTGGTCCATAAACAG	
3	Lb_ORF92_F	FORWARD	59.94	45	TGACCAAGACATGGTGGAAA	248
4	Lb_ORF92_R	REVERSE	60.07	45	CCGAATTGAATGACATGCTG	
5	Lb_ORF58_F	FORWARD	59.65	50	TACCAAATGGTGGAGGGAAC	250
6	Lb_ORF58R	REVERSE	59.60	40	CCATTTAAAACGTCGCAACA	
7	Lb_ORF68F	FORWARD	59.79	50	TGTCTGGAGATTGCCATCAG	239
8	Lb_ORF68R	REVERSE	60.04	45	CCAATTTTCGGAAGTGAGGA	
9	Lb_ORF5F	FORWARD	60.41	40	GATTCGCCAAATTTGATTGC	243
10	Lb_ORF5R	REVERSE	60.08	45	ATCATCATTGTCAGCGTCCA	
11	Lb_ORF60F	FORWARD	59.89	50	ACGTACGATTGGCGTAAACC	235
12	Lb_ORF60R	REVERSE	60.84	55	GACGTTGTTGTCCGAAGAGC	
13	Lb_ORF85F	FORWARD	59.77	40	CAGCTTTAGAACCTTGGGAAAA	249
14	Lb_ORF85R	REVERSE	59.73	45	GCCAACGCTGCACATTATTA	
15	Lb_ORF78F	FORWARD	60.07	45	CGATTTTGATGGTGATGCAG	251
16	Lb_ORF78R	REVERSE	59.31	40	CATTTTCAATGCACGAAAGC	
17	Lb_ORF94F	FORWARD	60.22	45	TGCCGTCGAAGATACATTCA	252
18	Lb_ORF94R	REVERSE	58.85	50	TCCACGGTAGACCATGTGTT	
19	Lb_ORF107F	FORWARD	59.62	55	CGACGCTATTGCAGTCAGTC	251
20	Lb_ORF107R	REVERSE	60.00	45	GCGTCAGAAGCAACAAATGA	
21	Lb_ORF87F	FORWARD	60.21	35	TTGCAATATGCCAGCAAAAA	260
22	Lb_ORF87R	REVERSE	59.92	40	GTTCCCAGGCAAAAATTTCA	
23	Lb_ORF72F	FORWARD	59.96	45	CTTTTTGCGGATCTTTCAGC	236
24	Lb_ORF72R	REVERSE	60.66	55	CTCCATTCTTGCCTGGACAC	
25	Lb_ORF83F	FORWARD	56.00	40	ATTCCAATGGTTGGCGAATA	84
26	Lb_ORF83R	REVERSE	62.00	55	CCGAGTGGAGTACACGTTTG	
27	Lb_RhoGapF	FORWARD	56.00	40	AATTCGGAAGCAATGGAAGA	325
28	Lb_RhoGapR	REVERSE	56.00	40	ATCGCTTGGTTTCTTTTTGC	
29	Lb_actinF	FORWARD	66.00	65	GATGCCCCGAGGCTCTCTTC	294
30	Lb_actinR	REVERSE	60.00	52	TGGTGCCAAGGCAGTGATT	
31	Lb_shakeF	FORWARD	64.00	60	CGAGTTATCGGTGCGCTTCC	182
32	Lb_shakeR	REVERSE	62.00	55	GCGAGGGACATCGCTTGATT	

Table S5: Primers used in the paper.

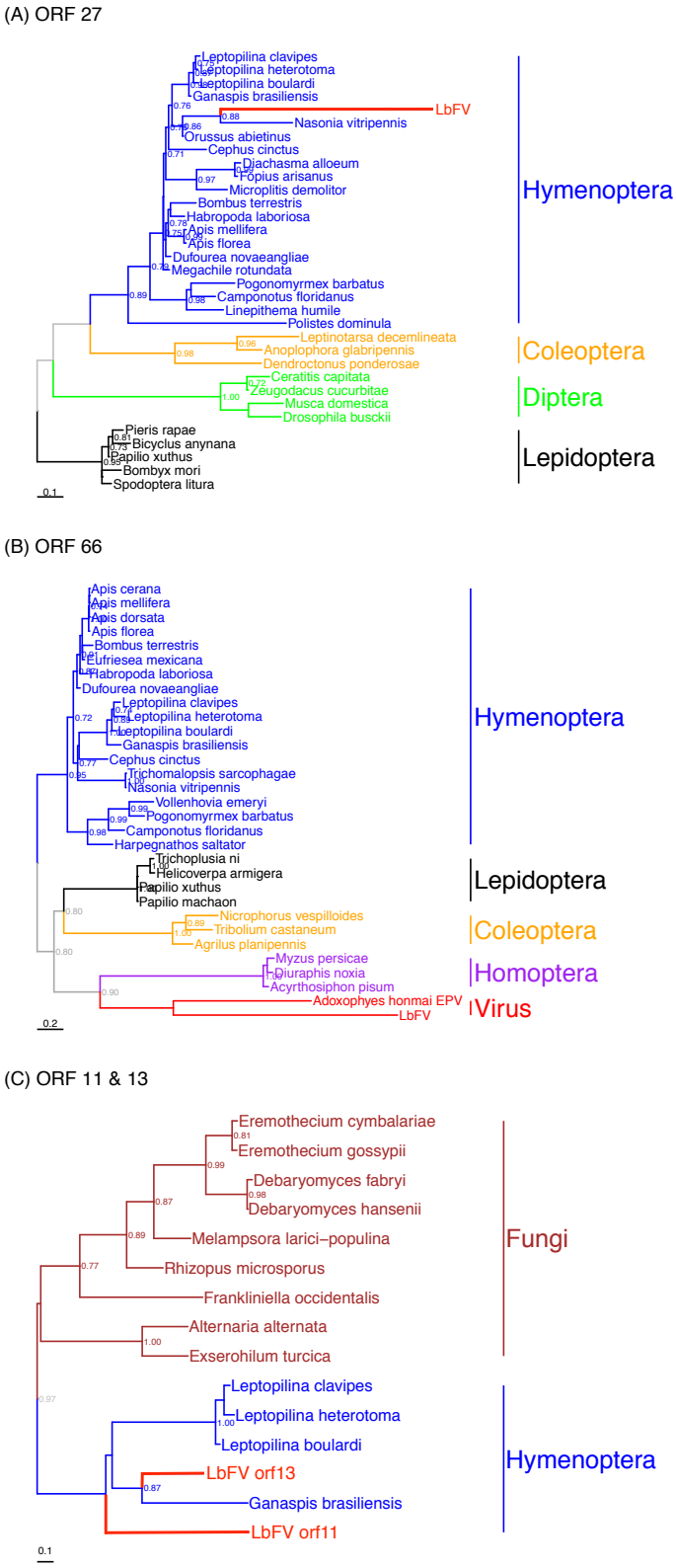

Figure S1: Four loci in the LbFV genome probably derive from insect genes. ORFs 27 (A) and 66 (B) are putative inhibitors of apoptosis and ORF 11 and 13 (C) contain a putative histone demethylase domain [64]. Sequences were aligned using muscle, and conserved blocks were identified using gblocks to construct a PhyML phylogeny (parameters: -d aa -m LG -b -4 -v e -c 4 -a e -f m). Only aLRT values > 0.7 are shown. Accession numbers of the corresponding sequences are available in table S4

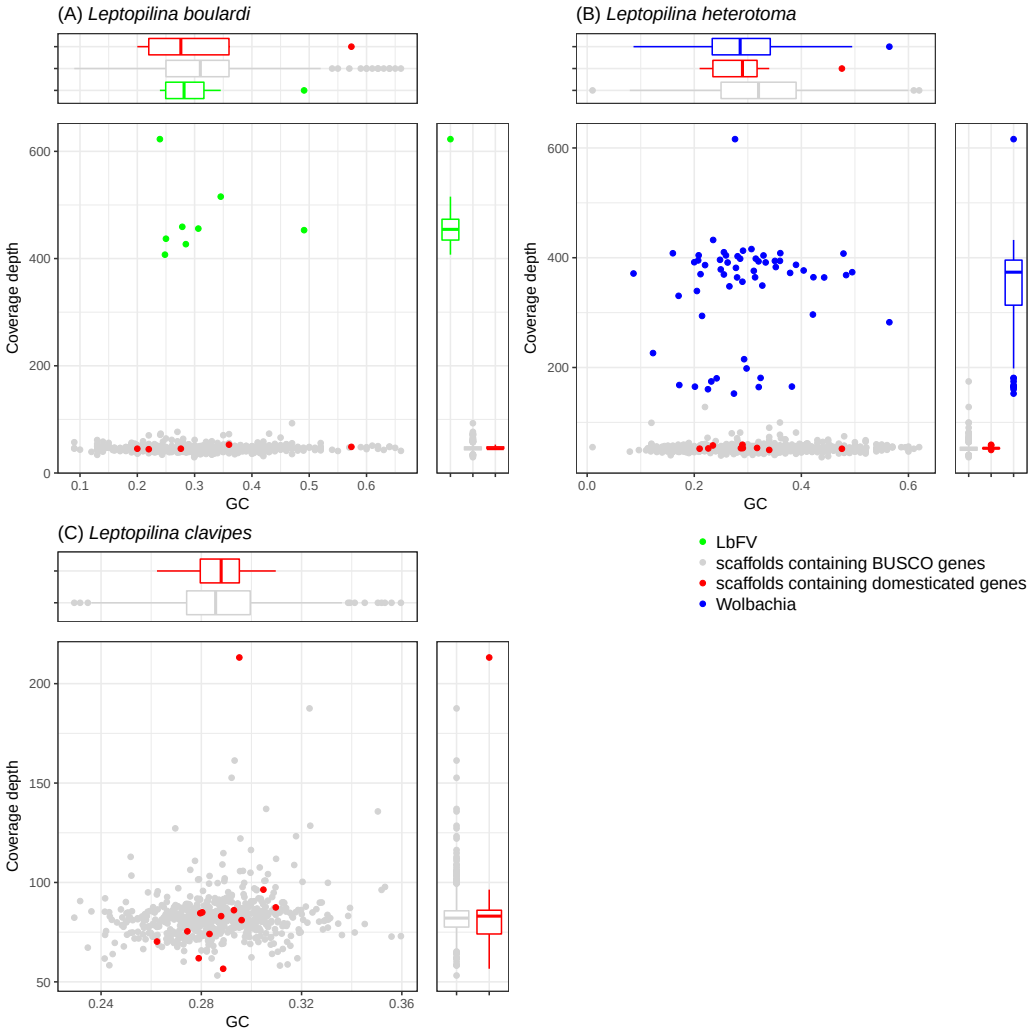

Figure S2: General features of scaffolds containing single copy universal arthropod genes (BUSCO gene set, in grey), scaffolds containing virally-derived loci (in red), scaffolds belonging to the virus LbFV (in green, only in *L. boulardi*) and of scaffolds belonging to *Wolbachia* endosymbiont (in blue, only in *L. heterotoma*). The heterogeneity in coverage depth for the *Wolbachia* scaffolds in *L. heterotoma* is probably the consequence of multi-infection with three *Wolbachia* strains having different densities[49]. (A) *L. boulardi*; (B) *L. heterotoma*, (C) *L. clavipes*.

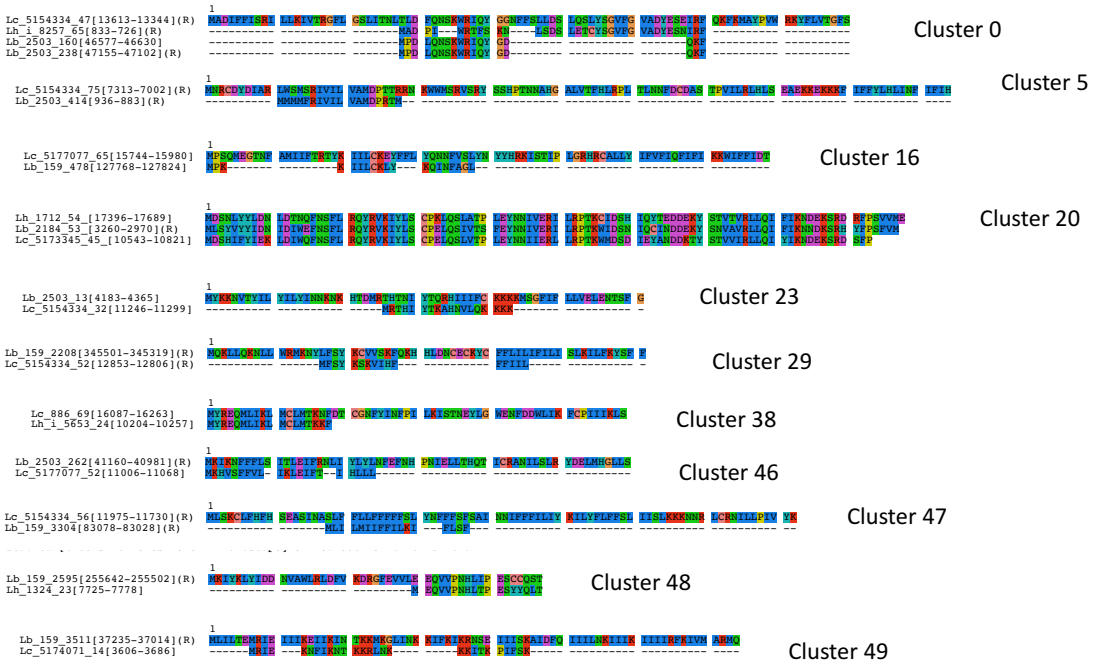

Figure S3: Flanking regions of virally-derived genes show similarities between *Leptopilina* species. Amino-acid sequences were predicted from the wasps scaffolds containing the virally-derived genes (but masked for the viral genes themselves) using getorf (-minsize 50 -find 1). They were clustered using CD-hit (-c 0.7), and aligned using muscle.

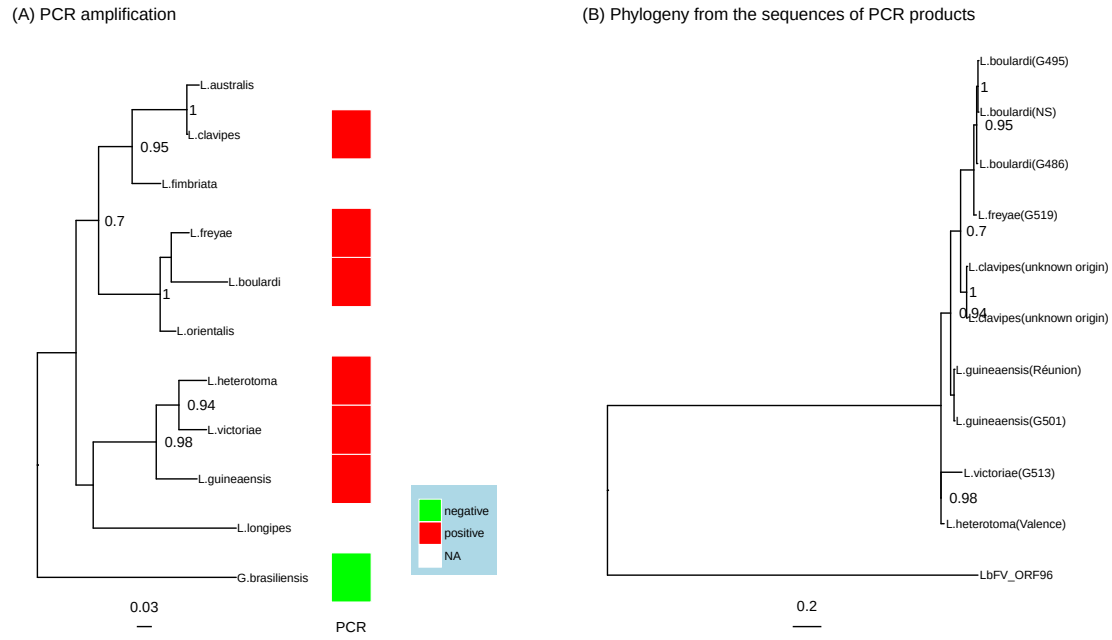

Figure S4: Amplification, sequencing and phylogeny of orthologs of LbFVORF96 in *Leptopilina* species. (A) Phylogeny of *Leptopilina* genus and *Ganaspis brasiliensis* based on internal transcribed spacer 2 (ITS2). (B) Phylogeny obtained after sequencing the corresponding PCR products. The stain used is indicated between brackets. Accession numbers of the corresponding sequences are available in table S4.

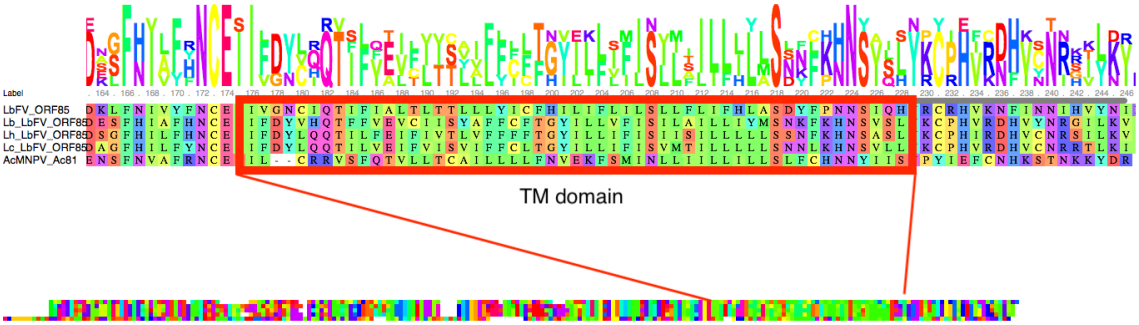

Figure S5: Ac81 homologs in LbFV and in *Leptopilina* genomes (ORF85) share a conserved hydrophobic, probably transmembrane domain.