Genome Sequencing of the Behavior Manipulating Virus LbFV Reveals a Possible New Virus Family

David Lepetit[1], Benjamin Gillet[2], Sandrine Hughes[2], Ken Kraaijeveld[3], and Julien Varaldi[1,*]

[1]Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, France

[2]Université de Lyon, CNRS, Ecole Normale Supérieure de Lyon, Université Lyon 1, Institut de Génomique Fonctionnelle de Lyon UMR 5242, France

[3]Department of Ecological Science, Faculty of Earth and Life Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

*Corresponding author: E-mail: julien.varaldi@univ-lyon1.fr

Accepted: November 16, 2016

Data deposition: All sequences produced for this article have been submitted to genbank Virus sequences: KY009685, Wasp sequences: KU948204.1 and KU948205.1 (*L. boulardi*), JUFY00000000.1 (*L. clavipes*), KU948207.1 and KU948206.1 (*L. heterotoma*), KU948208.1 (*Ganaspis* sp). Other publicly available sequences used in the article are listed in supplementary tables S4 and S5, Supplementary Material online.

Abstract

Parasites are sometimes able to manipulate the behavior of their hosts. However, the molecular cues underlying this phenomenon are poorly documented. We previously reported that the parasitoid wasp *Leptopilina boulardi* which develops from *Drosophila* larvae is often infected by an inherited DNA virus. In addition to being maternally transmitted, the virus benefits from horizontal transmission in superparasitized larvae (*Drosophila* that have been parasitized several times). Interestingly, the virus forces infected females to lay eggs in already parasitized larvae, thus increasing the chance of being horizontally transmitted. In a first step towards the identification of virus genes responsible for the behavioral manipulation, we present here the genome sequence of the virus, called LbFV. The sequencing revealed that its genome contains an homologous repeat sequence (*hrs*) found in eight regions in the genome. The presence of this hrs may explain the genomic plasticity that we observed for this genome. The genome of LbFV encodes 108 ORFs, most of them having no homologs in public databases. The virus is however related to Hytrosaviridae, although distantly. LbFV may thus represent a member of a new virus family. Several genes of LbFV were captured from eukaryotes, including two anti-apoptotic genes. More surprisingly, we found that LbFV captured from an ancestral wasp a protein with a Jumonji domain. This gene was afterwards duplicated in the virus genome. We hypothesized that this gene may be involved in manipulating the expression of wasp genes, and possibly in manipulating its behavior.

Key words: LbFV, phylogeny, horizontal gene transfer, manipulation, recombination, behavior.

Introduction

In host-symbiont interactions, selection may favor the evolution of parasitic strategies that favor the reproduction of the symbiont even if this comes at a cost on host fitness as far as host and symbiont genes are not strictly co-transmitted along generations (Smith 2007). This situation of evolutionary conflict is typical of host-parasite interactions. In particular, this conflict may explain why a wide diversity of parasites alter the behavior of their hosts, sometimes with spectacular extended-phenotypes (phenotype corresponding to the joint expression of the genes of the host and the parasite). For instance, many trophically transmitted parasites modify the behavior of their host in a way that enhances the chance of being predated, thereby increasing their chance to reach the next host (Moore 2013). Examples of such phenomena include mice infected by the protozoan *Toxoplasma gondii* that renders infected mice attracted by the odor of cats' feces, thus facilitating the transmission to its definitive host (McConkey et al. 2013). Striking examples of host behavior manipulation involve parasites as diverse as trematodes, bacteria, fungi and viruses, showing that evolution has repeatedly favored such manipulative strategy.

We previously described a case of behavior manipulation involving a virus and a parasitic wasp. In this system, the wasp

Leptopilina boulardi lays its eggs into *Drosophila* larvae. The wasp then develops within the *Drosophila* larva which continues its own development. However, once the *Drosophila* has pupated, the developing parasitoid starts to consume essential organs of the *Drosophila*, ultimately killing the fly. Parasitoid wasps, like most Hymenoptera, display sophisticated behavior (Godfray 1994). In particular, when foraging females encounter a new host, they have the capacity to detect whether this potential host has already been parasitized or not. In most conditions, parasitoid females refuse to lay eggs in already parasitized hosts. This decision makes sense since the second parasitoid egg is usually at competitive disadvantage in "superparasitized" hosts. However, females sometimes superparasitize. This paradoxical decision was first interpreted as being the consequence of "errors" on the part of the egg-laying females (Van Lenteren & Bakker 1975). Later, this interpretation was challenged by experiments showing that females superparasitize only when they have no better solution around, i.e. when the overall environment quality is low. This last observation led scientist to formulate an adaptive interpretation of this behavior: laying a second egg in a host may still pay if this egg has a nonnul probability to win the competition for the possession of the host, which is the case. Several experiments and theoretical models, in particular done using *Drosophila* parasitoids as model system, globally validated this interpretation (Van Alphen & Visser 1990).

Contrary to what was expected, we found that superparasitism behavior in *L. boulardi* is in fact mostly under the control of a DNA virus called LbFV for *Leptopilina boulardi* Filamentous Virus (Varaldi et al. 2003, 2006). In other words, LbFV forces the females to accept already parasitized hosts. Interestingly, in situations of superparasitism, the virus is able to be horizontally transmitted to the developing parasitoid. Thus, this behavior modification (induction of superparasitism), directly favors the horizontal transmission of the virus. Accordingly, by a theoretical approach, we showed that the virus is always selected to increase the natural tendency of the wasp to superparasitize, which indicates a conflict of interest on this trait, and shows that this is a case of behavior manipulation (Gandon et al. 2006). In addition to horizontal transmission under superparasitism, the virus is vertically transmitted (from mother to offspring) with very high efficiency and reaches high prevalence (~95%) in some natural populations (Patot et al. 2010). The effect of the virus is mostly restricted to superparasitism. For instance, females have similar lifetime expectancy, although they incur a modest cost on size (Varaldi et al. 2005). Interestingly, the virus also brings a slight protection for the wasp egg against the immune reaction of some *Drosophila* strains (Martinez et al. 2012), underlying the multidimensionalty of the relationship between the virus and the wasp.

Electron microscopy investigations in female wasp abdomen revealed that LbFV produces large (~1 μm) flexible rod-shaped enveloped particles. The replication of the virus within a virogenic stroma, typical of DNA viruses, has been observed in the nucleus of cells in wasp ovaries. We also found that both ovaries and poison gland extracts are infectious to uninfected parasitoid larvae and can durably change the behavior of adult wasps and their offspring (Varaldi et al. 2006). A previous transcriptomic work gave access to the first molecular sequence of LbFV, and confirmed that LbFV is a DNA virus (Patot et al. 2009).

In an attempt to better characterize this peculiar virus and to give the opportunities to identify the molecular determinants underlying the behavior manipulation induced by LbFV, we present here its genomic sequence and its genomic structure. This work revealed that LbFV is related to the Salivary Gland Hypertrophy Viruses found in the tstse fly and the domestic fly (Abd-Alla et al. 2009), but is still very distant to them, possibly belonging to a new virus family. We also found that LbFV captured several eukaryotic genes. The putative function and evolutionary history of these genes are discussed.

Material and Methods

Behavior Assay

To test the phenotype of females, we used the same protocol as in Varaldi et al. 2003. A single isolated female 1 to 2-day old was placed into a Petri dish containing ten *Drosophila* larvae (first instar) on a layer of agar and baker yeast. The female was introduced at 5 pm and removed the following day at 10 am (i.e. after 17h exposure). Two or three days after, a sample of three *Drosophila* larvae were dissected under the binocular microscope to count the number of parasitoid eggs and/or larvae. The behavior of each female was measured as the mean number of eggs per parasitized *Drosophila* larva. All experiments and rearings were conducted in a climatic chamber at 25 °C and 70% humidity. Parasitoid originated from Sienna (Italy) and the virus originated from Gotheron (near Valence, France). The virus was transferred from the Gotheron strain to the (initially uninfected) Sienna strain by natural superparasitism as described in (Varaldi et al. 2003). Parasitoids were reared on the same *Drosophila melanogaster* strain (originating from Sainte-Foy-lès-Lyon, near Lyon, France).

Isolation of Genomic DNA of LbFV

About 2 g of adult parasitoid wasps were homogenized in a buffer (20 mM Tris–HCl pH 7.5, 10 mM KCl, 4 mM $MgCl_2$, 6 mM NaCl, 1 mM Dithioerythrol, 0.1% Tween-20) using a Tissue-Lyser apparatus (Qiagen) during 30 s at 30 Hz. The solution with crushed wasps was centrifuged at 1,000×*g* during 5 min and the supernatant was filtrated through three successively filters (hydrophilic, Minisart, Sartorius) with

pore size decreasing from 5 μm to 1.2 μm and to 0.45 μm. Final concentrations of 400 μg/ml RNaseA and of 0.5 units/μl DNaseI were adjusted in the flow-through and incubated 2 h at room temperature. Nucleic acids were then extracted with SDS-proteinase-K and phenol-chloroform-pH7 and were precipitated with isopropanol. The pellet of nucleic acids was resuspended in TE 10-1 with RNaseA and nucleic acids of low molecular weight were removed with a precipitation in 6.5% PEG 6000 and 0.8% NaCl. Purified nucleic acids were digested by DNase I and RNase A to determine their nature. Digested nucleic acids were then treated with phenol/chloroform/isoamyl alcohol (25:24:1, pH 6.8) and precipitated with ethanol and sodium acetate before resolution on agarose gel.

Sequencing Strategy

The purified DNA was first sequenced using the 454 technology on a GS Junior platform. Library construction was performed with the Roche dedicated kit (Lib L protocol) according to manufacturer instructions. About 117,777 reads were obtained with a mean read length of ~370 bp. Because of the high AT content of the genome (~80%), the genome draft contained long homopolymers and the 454 technology gave numerous sequencing errors. We thus decided to sequence again on an Illumina Miseq platform. The library was prepared using the Ovation Ultralow kit (Nugen) starting from ~40 ng of purified virus. The mean insert size was ~580 bp (min~450 bp max~850 bp) and read length was 2×250 bp. About 14,474,973 paired-end reads of high quality (mean quality score 34) were obtained.

Long-Reads Sequencing

Because the previous sequencing data did not lead to a single circular chromosome (as was expected from preliminary data), we decided to generate long reads using MinION Nanopore sequencing (Oxford Nanopore Technologies, ONT). The quantity of DNA extract (~40 ng) being limited, the Low Input protocol for genomic DNA was used to construct the library, following ONT and manufacturers' instructions. The DNA was first sheared with a Covaris g-TUBE leading to fragments distributed ~8 kb. After end-repair and dA-tailing performed with New England Biolabs (NEB) reagents, the Low Input Expansion Pack was used in conjunction with Nanopore sequencing kit (ONT SQK-MAP006) to build the final library. The totality of the library, without supplementary purification step, was deposited on a flow cell (R9) connected to a MinION MK1.0. The 48-h genomic DNA sequencing script was run in MinKNOW v0.51.3.40, but stopped after 44 h, and the Metrichor v2.39.3 was used for base calling. Only 2D reads (7,996 reads generated) were taken into account for the further analyses and extracted using the minoTour interface (http://minotour.nottingham.ac.uk/) developed by Matthew Loose (Loose 2014).

Assembly Strategy for 454 + Illumina Dataset

After quality trimming, the 454 reads were assembled using Newbler 2.5.3 using default parameters. This led to ten putative viral contigs from 1,559 to 24,643 bp.

From the 14 million of paired-end Illumina reads, we took only 100,000 quality-trimmed reads to feed the Velvet assembler (version 1.2.09). We used only 100,000 reads (corresponding to a final 170× coverage) because we observed that Velvet gave not meaningful results with higher number of reads. We tested various kmer values (from 25 to 101). Since we did not have any reference genome, it was difficult to estimate the reliability of assemblies based on simple statistics like N50 because misassemblies may artificially increase N50. Thus, we used REAPR (Hunt et al. 2013) which is designed to use mapping of paired-end reads to evaluate the reliability of an assembly. Contigs are then cutted at points where coverage data suggest misassemblies. The software produces corrected N50 and other statistics calculated after this contigs-breaking process. Based on that and on further visual inspection of mappings, we choose k57 as the best assembly. Because REAPR suggested misassemblies for all k values, including the k57 assembly, we choose to run another assembly without scaffolding with $k = 57$.

To confirm the sequence of these contigs, we used the software Price (Ruby et al. 2013). This software performs a paired-read iterative contig extension and is designed to assemble sequences starting from a seed. Starting from 500 bp seeds defined in the middle of each viral contigs of the assembly k57 (without scaffolding), Price was used to extend the contigs. Price confirmed the sequence previously obtained with Velvet and extended all but one contig of 10 bp to a few kb. We ended up with eight putative viral contigs. REAPR was run again using 240,000 independent reads on this new assembly and the results suggested that one contig should be split in three parts. Accordingly, Price did not extend this contig but instead gave three sub-parts starting from three different seeds (one in the 5′ part, one in the middle, one in the 3′ part). Accordingly, the contig "70" has been split in three parts.

Finally, we compared this assembly with the assembly previously obtained with 454 reads using the software Mauve version 2.3.1 (Darling et al. 2004). The draft obtained with the 454 reads indicated two contig connexions that were not found in the Miseq genome draft.

Before the addition of the long-reads, the LbFV draft contained eight contigs from 2,213 to 24,608 bp and was further validated by running the REAPR pipeline. Most of the sequence manipulation was done in R using the package seqinr (Charif & Lobry 2007). The data for the construction of the De Bruijn graph were extracted from the LastGraph file produced by Velvet and visualized by Cytoscape (Shannon et al. 2003).

Identification of Repeated Sequences

To identify large homologous sequences (*hrs*), we used a reciprocal BLAST approach. The draft was blasted against itself using default parameters, and it was evident from this analysis that a large repeated sequence was present at each extremities of the contigs (obtained using the 454 and Illumina sequencing.) We then extracted the 2,500-bp left and right flanking regions of each contig and aligned them after reverse complementation for right extremities. The 16 sequences were aligned using MUSCLE (Edgar 2004) as implemented in SeaView (Gouy et al. 2010). By eye, a subset of ten sequences was selected because they appear to align on a larger portion of the sequence. Based on these ten aligned sequences, a consensus sequence was derived (60% threshold). Finally a block of highly conserved 425 bp was defined by visual inspection. This 425bp consensus sequence was then used as a query in a BLAST against the genome of LbFV to identify all full or partial copies of this repeated element. All hits were then extracted and aligned using MUSCLE and a 449bp consensus sequence was derived in Jalview (http://www.jalview.org/). The secondary structure of this sequence was predicted using the software RNAstructure with default parameters except that we selected the DNA option.

(http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html).

Simple direct repeats were searched for using Tandem Repeats finder (Benson 1999) ran on a MacOS X with default parameters. Only simple repeats with score > 100 were considered in the analysis and represented in the figure. Inverted repeats were searched for using the online software einverted (http://emboss.bioinformatics.nl).

A draft of the LbFV genome of LbFV masked for hrs sequence and any repetition was obtained by a reciprocal BLAST approach. Sequences with percentage identity >90% and with >60 nucleotides aligned were replaced by Ns.

Combinatory PCR

To identify the connexions between the eight contigs (obtained after the 454 and Illumina sequencing), we designed 16 highly specific primers in nonrepeated flanking regions of each contig (Tm ~70 °C, see table S1). The 3′ end of the primers was oriented towards the nearest end of the contig (either left or right side) in order to amplify the putative DNA present between the contigs. All combination of pairs of primers were included in the PCR assay, including controls without DNA or with a single primer. PCRs were performed in 20 µl containing 1 µl of viral nucleic acid template (similar to the DNA extract used for sequencing), 1× Phire Reaction Buffer, each dNTP at a concentration of 50 µM, each of the two primers at a concentration of 0.5 µM, 3% DMSO and 0.4 µl of Phire Hot Start II DNA Polymerase (Thermo Scientific). The following cycling program was used: Initial denaturation 30 s at 98 °C, Cycling conditions: 5 s at 98 °C, 5 s at 68 °C, and 2 min at 72 °C, final extension 5 min at 72 °C (Tetrad2 DNA engine, BioRad).

Scaffolding Using MinION Long-Reads

The MinION run produced 7,996 2D reads with a mean length of 4,183 bp (3d quartile 6,045 bp). Reads with length >3 kb were blasted against the eight contigs previously identified (masked for the hrs sequence) using high stringency criteria (e-value < 10e-20 and alignment length > 1,000 bp). The mean identity between MinION reads and the previously obtained contigs (masked for repeated regions) was 78% which is consistent with the known error rate of MinION technology (Jain et al. 2015). The reads mapping to two contigs were then collected and aligned with the ends of the two corresponding contigs using mafft v7.294b (default parameters, Katoh et al. 2002). This allowed to generate a circular molecule of 111,453 bp.

Validation of the Assembly

To test the validity of the final assembly, we mapped 100,000 MiSeq reads and the totality of 2D MinION reads ($n = 7,996$) against the final circular draft genome. Miseq reads were mapped using bowtie2 with default parameters-except -X 1200 (Langmead & Salzberg 2012). Discordant mapping were removed using samtools (Li et al. 2009). MinION 2D reads were mapped using the software graphmap using the following parameters: -C -t 8 -B 50 (Sovic et al. 2016). By eye inspection of the mapping of Illumina reads (that have low error rates), we corrected a few small residual errors present in the draft (1 bp errors and 1 bp insertions/deletions). After a round of correction, the reads (both Illumina and MinION reads) were mapped again to verify the quality of the corrected draft. The procedure was repeated 4 times until achieving a satisfying result (111,453 bp). Note that since MinION has a high error rate, we decided to replace by Ns the sequences that were obtained only with MinION reads (hrs-containing regions) in the released genome. Nevertheless, in this article we also analyzed the hrs-containing regions by studying the consensus MinION sequences. In particular, we tested for the presence of conserved motifs in these hrs-containing regions using the online MEME server (Bailey et al. 2009) with the following parameters -dna -oc -nostatus -time 18,000 -maxsize 60,000 -mod anr -nmotifs 8 -minw 6 -maxw 50 -revcomp. We also studied the predicted 2D structure of these regions by using the RNAfold webserver with default parameters (except DNA option).

Gene Content and Gene Phylogenetic Reconstruction

Gene prediction was performed using the ORFFinder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Only open reading frames (ORFs) starting with a methionine and ending with a stop codon, with at least 50 amino acids and

with minimal overlap (<23 nucleotides) were considered as valuable candidates for being true ORFs.

The 108 predicted ORFs were analyzed against Uniprot using the hmmer webserver (Finn et al. 2011) using a threshold of 0.001 and a BLOSUM45 substitution matrix which is more suited to divergent alignments. For each protein, we ran the phmmer for searching similarities with known proteins and the hmmscan for the identification of conserved domains. Both algorithms were ran with default parameters. Only domains with individual e-value < 0.0001 were reported. For ORFs with eukaryotic homologs in public databases, we additionally searched for viral homologs in the draft genomes of *Leptopilina* species (*L. boulardi*, *L. heterotoma*, *L. clavipes*) and the related Ganaspis sp. (unpublished). Proteins were predicted manually from the BLAST output (for *Leptopilina* species) or with Augustus gene prediction webserver using Nasonia as gene model (for *Ganaspis* sp.).

To construct phylogenies, we selected a representative subset of the proteins identified by phmmer, aligned them using muscle and selected conserved blocks using Gblocks as implemented in the command line version of seaview (Gblocks was run with options -b2 -b4 -b5). The phylogenies were constructed using PhyML (with options -d aa -m LG -b -4 -v 0.0 -c 4 -a e -f m). The branch supports were estimated by approximate Likelihood ratio tests (aLRT).

Promoter Analysis

We searched for conserved DNA motifs in the 300bp upstream of each predicted ORF using the webserver MEME (Bailey et al. 2009) with the following parameters -dna -oc -nostatus -time 18,000 -maxsize 60,000 -mod anr -nmotifs 8 -minw 6 -maxw 50 -revcomp. The first identified motif corresponded in fact to a portion of the identified *hrs* sequence. This motif was not represented in the figure (supplementary fig. S4, Supplementary Material online). Logo motifs were then obtained using the webserver weblogo (Crooks et al. 2004).

Phylogenomic Analysis

In order to position LbFV in the phylogeny of large dsDNA viruses of arthropods, we constructed a local database containing the predicted proteins from 13 arthropod dsDNA viruses (see table 1). We then blasted the predicted proteins from LbFV on this database with an e-value cutoff of 0.01. From this analysis we identified six LbFV ORFs that had at least four hits with the 13 virus species present in the database (ORF37, ORF52, ORF58, ORF85 and ORF106 see supplementary table S5, Supplementary Material online). The proteins were aligned using MUSCLE and conserved blocks were identified using Gblocks as implemented in the command line version of Seaview (Gblocks was run with options -b2 -b4 -b5). The alignment of conserved blocks were concatenated and a phylogeny was constructed on the combined alignment using PhyML (with options -d aa -m LG -b -4 -v 0.0 -c 4 -a e -f m). The branch supports were estimated by approximate Likelihood ratio tests (aLRT). The congruency of the phylogenetic signal among the six genes was tested by comparing the likelihood of individual gene tree (unconstrained) to the likelihood of a constrained tree using the final tree as the constraint. This was performed using RaxML (with a LG substitution matrix) with an SH-test (Stamatakis et al. 2012). All six genes showed congruent phylogenetic signal with the final phylogeny based on concatenated data. Preliminary phylogenetic analysis suggested that these ORFs were not acquired by horizontal transfer but were rather vertically transmitted since the divergence with other large dsDNA viruses.

Results

Based on previous results it was known that the virus LbFV was responsible for the superparasitism behavior in *L. boulardi*. We established two lines with the same genetic background (origin Sienna, Italy), but displaying very contrasting behavior (Varaldi et al. 2006). The virus-uninfected NS ("nonsuperparasitizing") line almost systematically refuses superparasitism (fig. 1*A*) whereas the S ("superparasitizing") line, infected by the virus LbFV, displays intense superparasitism tendency (fig. 1*B*). Importantly, both lines lay only a single egg per oviposition but sharply differ in their decision to accept already parasitized hosts. The S line (Sienna) was used to further characterize the virus.

Preliminary Analysis of the Structure of LbFV Genome

Following extraction and purification of the viral genetic material (from the S line), a gel electrophoresis revealed two bands, one of them with a high molecular weight (fig. 2*A*). This band was absent from nonsuperparasitizing lines (not shown) and was DNAse sensitive but not RNAse sensitive (fig. 2*B*). The smaller band present in both lines was in fact an RNA virus related to Totiviridae and has been described elsewhere (Martinez et al. 2015). We were not able to visualize a clear band after pulsed-field gel electrophoresis (not shown). The quantity of viral DNA obtained was low (~40 ng) even if we pooled thousands of wasps (~2 g) for a single purification. The DNA was first subjected to sequencing technologies available at that time and able to work with such low DNA quantities (454 and Illumina paired-end technologies). The reads obtained were assembled (see "Methods" section for details). Because preliminary morphological and genomic analysis of LbFV genome suggested that LbFV was related to other large dsDNA arthropod viruses (Baculoviruses, Nudiviruses, Hytrosaviruses), whose genome consists of a large circular dsDNA molecule, we expected to obtain a single contig after the assembly. However, independently of the sequencing technology (454/Illumina), and of the assembler used (Newbler, Velvet, Mira), we systematically ended up with several contigs presumably belonging to LbFV genome.

Table 1
Genomic features of representative dsDNA insect viruses

Virus name	Family	Accession no.	Genome size (bp)	*n* genes	AT	Coding density (%)	# BLASTp hits with LbFV ORFs as queries*	# best BLASTp hit with LbFV ORFs as queries*
*Lymantria dispar multiple*NPV	Baculoviridae	NC_001973	161,046	164	42.5	87.5	8	2
*Cydia pomonella*GV	Baculoviridae	NC_002816	123,500	143	54.7	90.1	7	2
*Autographa californica*NPV	Baculoviridae	NC_001623	133,894	156	59.3	97.2	5	0
*Neodiprion setifer*NPV	Baculoviridae	NC_005905	86,462	90	66.2	84.5	7	0
*Culex nigripalpus*NPV	Baculoviridae	NC_003084	108,252	109	49.1	91.2	4	0
*Glossina pallidipes*SGHV	Hytrosaviridae	NC_010356	190,032	160	72.0	86.5	16	9
*Musca domestica*SGHV	Hytrosaviridae	EU522111	124,279	108	56.5	90.9	20	8
*Gryllus bimaculatus*NV	Nudiviridae	EF203088	96,944	98	72.0	93.6	6	2
*Tipula oleracea*NV	Nudiviridae	KM610234	145,704	131	74.5	85.7	5	2
*Heliothis zea*NV-1	Nudiviridae	AF451898	228,089	154	58.2	69.4	8	2
*Oryctes rhinoceros*NV	Nudiviridae	EU747721	127,615	139	58.4	88.5	6	1
*Penaeus monodon*NV	Nudiviridae	KJ184318	119,638	115	65.5	95.6	6	0
*Apis mellifera*FV	unassigned	KR819915	496,396	247	49.2	65.0	12	4
***Leptopilina boulardi*FV**	**unassigned**	**KY009685**	**111,453**	**108**	**78.7**	**80.0**	–	–

*Represents e-values <0.01

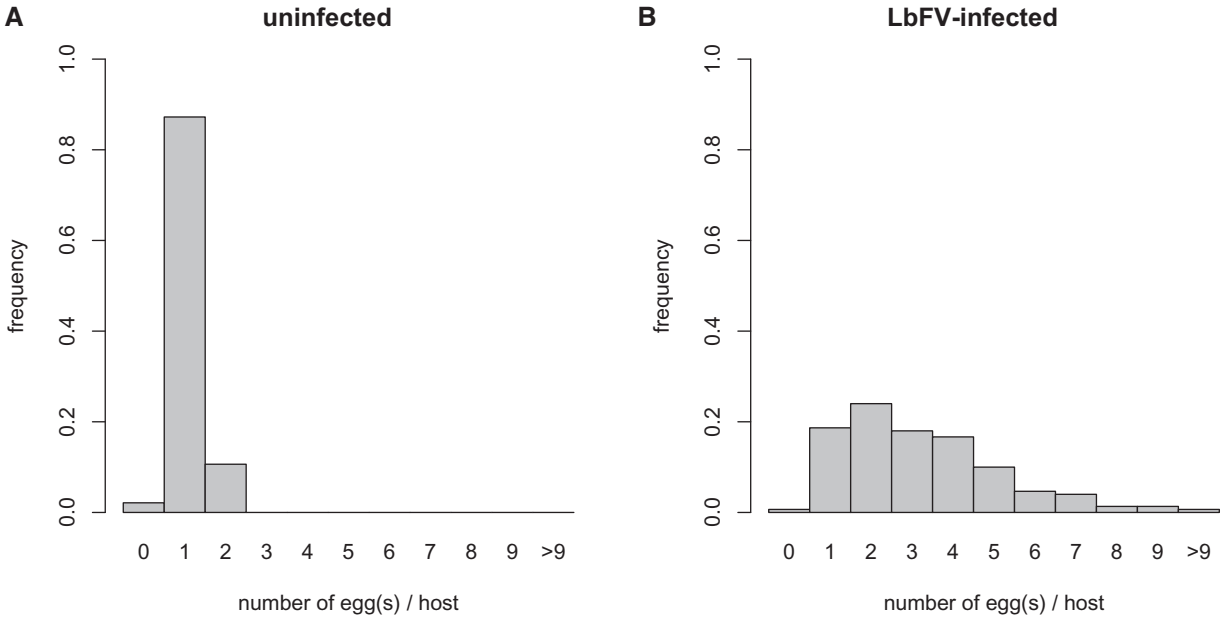

Fig. 1.—Superparasitism phenotype associated with LbFV infection: distribution of wasp eggs deposited by (*A*) nonsuperparasitizing uninfected females and by (*B*) LbFV-infected superparasitizing females within *Drosophila* larvae. Wasp were tested under controlled laboratory conditions (see "Methods" section for details). The number of female tested was 6 and 20, respectively, for NS and S lines.

After assembly and the subsequent bioinformatic pipeline, we obtained eight large contigs sharing similar GC content (~21.5%) and similar coverage (~170×). Most of these contigs (7/8) shared similarities with known viruses based on blast analysis (see below). Importantly, the sequences contained in these eight contigs were connected in the de Bruijjn graph generated by the assembler Velvet (fig. 3), suggesting that these eight contigs are part of a single DNA molecule. These eight contigs sum up to 113,482 bp which is in the range of other arthropods dsDNA viruses genome size.

Viral Contigs Are Flanked by Homologous Regions

We first searched for homologous regions (*hrs*) within or among those eight contigs using reciprocal BLAST. This

analysis revealed that all eight contigs contain one or several copies of a very similar sequence of ~449 bp. This sequence was only found in the 5′ and 3′ extremities of each contig and the different copies of it were highly similar with each other, with pairwise identity ranging from 85% to 100% and with a mean of 95% (fig. 4). All eight contigs were flanked with this homologous sequence either complete or partial. For some contigs, several partial or complete copies of this homologous sequence were present in the flanking regions. We found that the consensus sequence of all copies of this element is predicted to contain a peculiar secondary structure with an hairpin at its 3′ end (between positions 365 and 449) (supplementary fig. S1, Supplementary Material online). No other inverted sequence were identified by the software einverted.

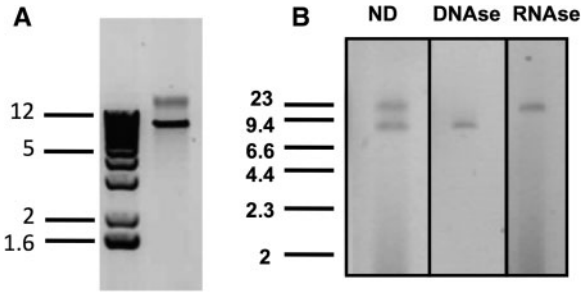

Fig. 2.—Characterization of the viral nucleic acids. (*A*) Electrophoresis of viral nucleic acids extracted from *Leptopilina boulardi* wasps and (*B*) viral nucleic acids were treated with DNAse I or RNAse A. ND, not digested. The ladders are expressed in kilobases.

Finishing the LbFV Genome

As previously mentioned, based on preliminary data showing its apparent relatedness with other large dsDNA viruses of insects, we expected the genome of LbFV to be composed of a single circular DNA molecule. Yet, in spite of numerous attempts to reduce this number either bioinformatically or experimentally, we systematically ended up with eight contigs flanked by this homologous region. Notably, similar assembly problems were

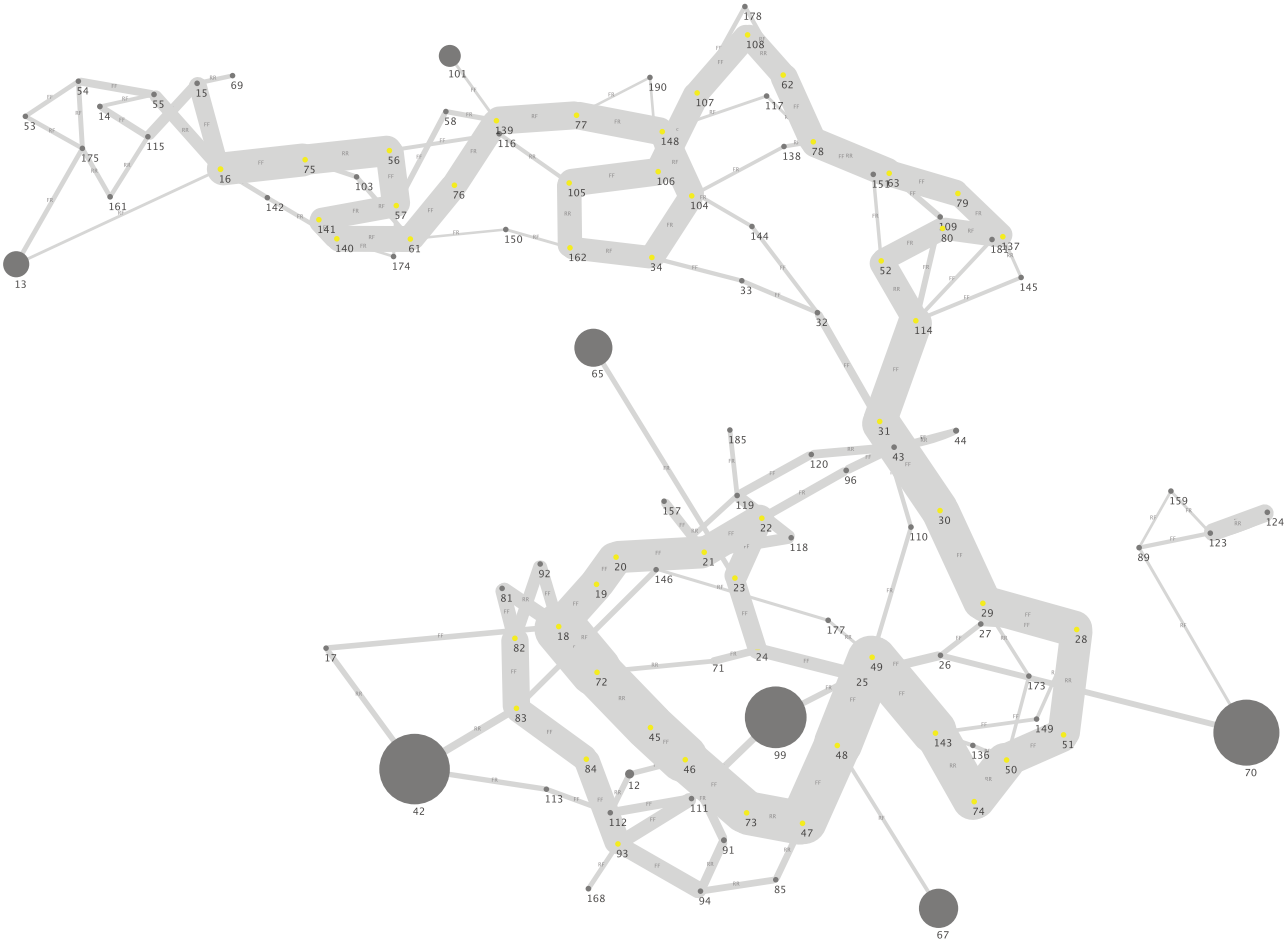

Fig. 3.—Simplified De Bruijn graph showing the connexions among viral contigs after assembly of Illumina reads. Each circle, so called node in the Velvet terminology, represents a contig. Its diameter is proportional to its length. Each edge connection between two contigs indicates contiguity among contigs. Edge width is proportional to the number of connexions found among nodes. The dark grey circles represent the eight large contigs identified by Velvet assembler (13, 101, 65, 42, 99, 70, 67, 12).

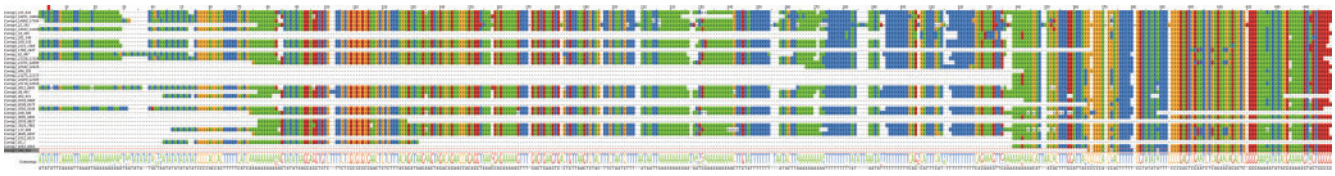

Fig. 4.—LbFV contigs identified after assembling Illumina and 454 reads are flanked by homologous regions (hrs). Alignment of homologous regions found in the LbFV genome. The contig name, start and end position of each sequence is indicated on the left. The last line of the alignment corresponds to the final consensus sequence of all contigs.

encountered when we sequenced the whole DNA extracted from a single LbFV-infected female (not shown).

To gain insights on the structure of LbFV genome, we blasted the 14 millions Miseq paired-end reads against the eight LbFV contigs masked for the *hrs* sequences and all other repetitive sequences in the genome. A BLASTn with default parameters was used. We then filtered out the pairs of reads that mapped to different LbFV contigs because they may reveal contig connexions. Only the reads that mapped with high specificity (percent identity $> 95\%$, alignment length > 60 bp and e-value $< 10^{-15}$) within the 2,000bp extremities of the contigs were considered. A total of 96 paired-end reads passed the filtering process. Various connexions were identified for each contig extremities (fig. 5*A*). This result suggests that LbFV has a certain recombinogenic ability possibly because of the presence of the large homologous regions *hrs*.

Next we designed PCR primers in nonrepeated sequences at both ends of each contig in order to amplify the putative gaps among the eight contigs. Among the 120 PCR (16 primers, all potential connexions investigated), 51 were positive. Importantly, most individual primer did amplify with several other alternative primers, suggesting that each contig has several connexions with other contigs (fig. 5*B*). Importantly, within a single PCR reaction, some primer combinations gave several bands within a single PCR reaction suggesting that polymorphism was also present in these peculiar genomic regions (fig. 5*C*). Accordingly, our attempts to Sanger sequence these PCR products always gave multi-peak sequences, suggesting the presence of polymorphism in these regions (not shown). This PCR-based assay suggested that the population of the virus is composed of several genome variants with the eight contigs showing different orientation and with polymorphism in between two consecutive contigs. However, we must stress that since those results are based on PCR amplifications, they do not inform on the relative abundance of each variant (neither on the arrangement among contigs nor on the nucleotidic variation between contigs).

Resolution of the Virus Genome Structure Using Long-Read Sequencing

In order to resolve the structure of the genome, we sequenced the same DNA extract by MinION Nanopore technology that produces long reads. This approach is expected to overcome the problem of repeat regions to assemble the eight contigs because the reads generated are long (many kbs), largely exceeding the size of the *hrs* motif (449 bp). We obtained 7,996 2D reads with a mean length of 4,183 bp (3d quartile 6,045 bp). Reads with length >3 kb were blasted against the eight contigs previously identified (masked for the hrs sequence) using high stringency criteria (e-value $<$ 10e-20 and alignment length $>$ 1,000 bp). The mean identity between MinION reads and the previously obtained contigs (masked for repeated regions) was 78% which is consistent with the high error rate of MinION technology. Still, the long MinION reads allowed us to scaffold the contigs. We selected the reads that blasted with two contigs ($n = 209$ reads), aligned them with the corresponding contigs and checked visually all the putative connexions. We found that these connexions defined a circular molecule encompassing the eight previously identified contigs. The order was the following: contig 1, contig 3 (reverse complement), contig 8 (reverse complement), contig 4, contig 6, contig 5, contig 2 (reverse complement), contig 7 (reverse complement) and again contig 1.

The number of MinION reads bridging the gaps between contigs was between 7 and 40 per connexion (the alignment files of each connexion are available upon request). Thus, the MinION data allowed us to scaffold the eight contigs previously obtained by 454 and Illumina sequencing leading to a single circular genome of 111,453 bp as was expected based on the relatedness of LbFV with other large dsDNA viruses of insects. Interestingly, the same order and orientation of contigs was obtained by assembling the MinION reads using the long-reads assembler CANU (Koren et al. 2016, result not shown). To further test the validity of the resulting draft, we mapped a sample of the Illumina reads ($n = 100{,}000$) and the totality of MinION 2D reads on the circular genome obtained (fig. 6). The data indicated that the coverage of Illumina reads was globally homogeneous but felt to zero between the contigs, as expected. On the contrary, the coverage of the long-reads MinION was homogeneous all along the circular genome indicating that the assembly is correct. Thus the MinION long-read sequences allowed to connect contigs and resolve the overall structure of the genome. However, because of the high error rate of MinION reads, we let the sequences connecting the eight contigs as unresolved (Ns)

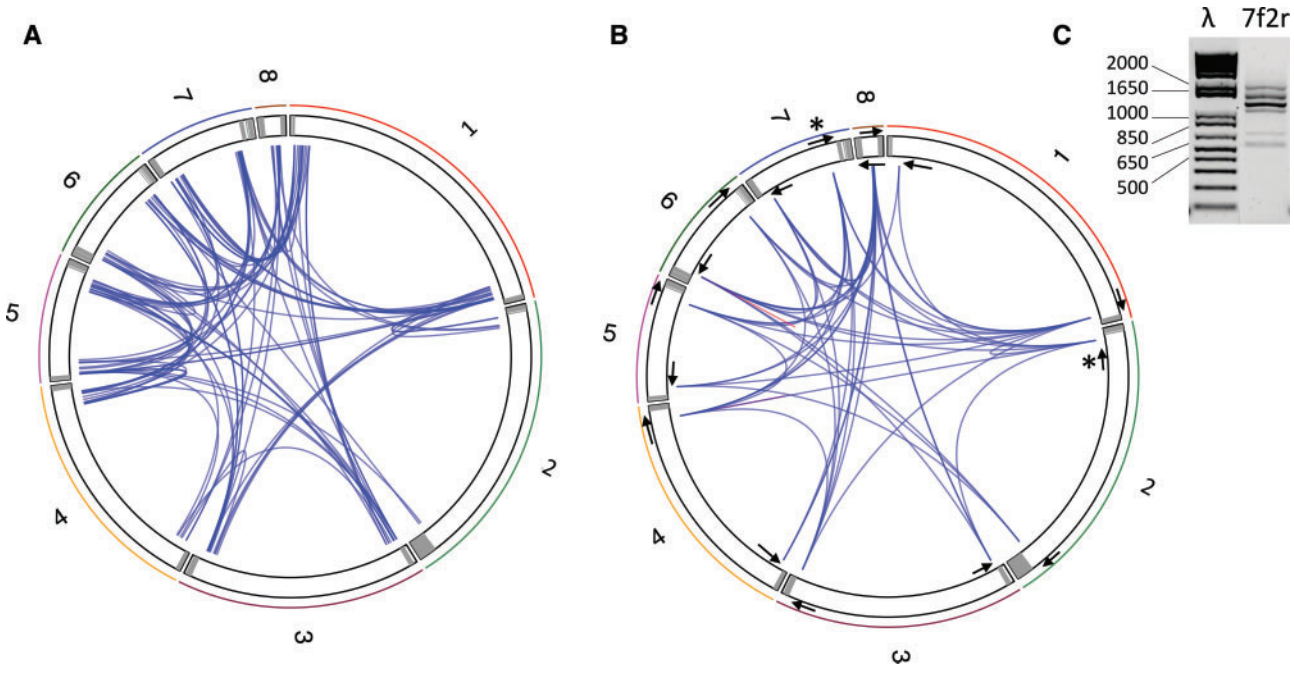

Fig. 5.—Connexion among contigs identified by (*A*) mapping paired-end reads on the LbFV contigs masked for repeated sequence. The eight LbFV contigs are arbitrarily ranked according to their length and represented within a circular chromosome. Grey areas indicate genomic regions homologous to the *hrs* motif. The blue lines represent the mapping of paired-end reads on different contigs extremities. In total 96 blue lines are represented corresponding to 96 paired Illumina reads. (*B*) PCR amplification. The 16 PCR primers used in this experiment are symbolically represented by the arrows (not scale). 51 out of 120 primer combinations gave a positive PCR and are represented by a blue line connecting contig extremities. The red lines indicate two cases where the PCR was positive with a single primer. (C) The PCR amplifications often gave several bands. As an example, the PCR product obtained with the 7f and 2r primers (identified in fig. 5*B* by stars) is shown. The full set of gel images is provided in supplementary figure S2, Supplementary Material online. λ, molecular weight marker; 7f2r, PCR product.

because only MinION reads are available in these genomic regions. Notably, we found no evidence of shuffled contig orientation in the MinION dataset, contrary as was found using PCR and Illumina data.

Analysis of the Complete Genome of LbFV: Repeated Elements

Because of the high error rate of MinION reads we let the sequence connecting two contigs as unresolved (Ns) in the fasta file submitted to genbank, since they are not reliable with ~22% error rate. However, we used the consensus sequence of those MinION reads to complement the data and to identify the general features of these eight hrs-containing genomic regions (see fig. 6 for the localization of the hrs-containing regions). The GC content of these regions was much higher than in other regions of the genome (0.347 vs. 0.213, χ^2 =251.12, df = 1, P value < 2.2 × 10e-16, fig. 6). Furthermore we identified eight highly conserved homologous blocks among the eight regions (supplementary table S2 and fig. S3, Supplementary Material online). All blocks were repeated several times within each genomic region with a clear nonrandom distribution, suggesting a function role for them. Moreover, the predicted 2D structure of those regions was extremely complex and stable, explaining the difficulty for obtaining good PCR amplification (supplementary fig. S5, Supplementary Material online). Several DNA viruses such as baculoviruses contain homologous regions in multiple locations of the genome. They function as origin of replication and for some baculoviruses also as enhancer of the transcription of adjacent genes (Hilton and Winstanley 2008).

In addition, direct repeats were also found in four putative gene coding sequences (ORF51, 61, 74, 108). ORF51 contains two repetitions of a 33bp motif (94% identity) followed by two repetitions of a 69bp motif (98% identity). Consequently, the predicted protein contains two peptide repetitions (ELGDKMPCKRK–ELGKKMPCKRK, KPSSSKIPNEK KDIINNDNNDDN-KPSSSKIPNEKKDIINNDNNDDN). ORF61 contains a direct repeat of 33 bp repeated 23.5 times (93% mean identity). Again, the predicted protein from this ORF contains a corresponding 23 repetition of the consensus peptide TTTTT STTLKP. ORF74 contains a 27bp motif repeated 3.3 times (96% mean identity) leading to the repetition of a 9 amino acid motif (ARTASPRKR) in the predicted protein. Finally, ORF108 contains a 21bp motif repeated 12.7 times that translate into a repetition of a 7amino acids peptide (EDKKIM). ORFs 46 and 47 correspond to the direct repetition of a 194bp DNA motif. In addition, ORF33 and ORF35 were very

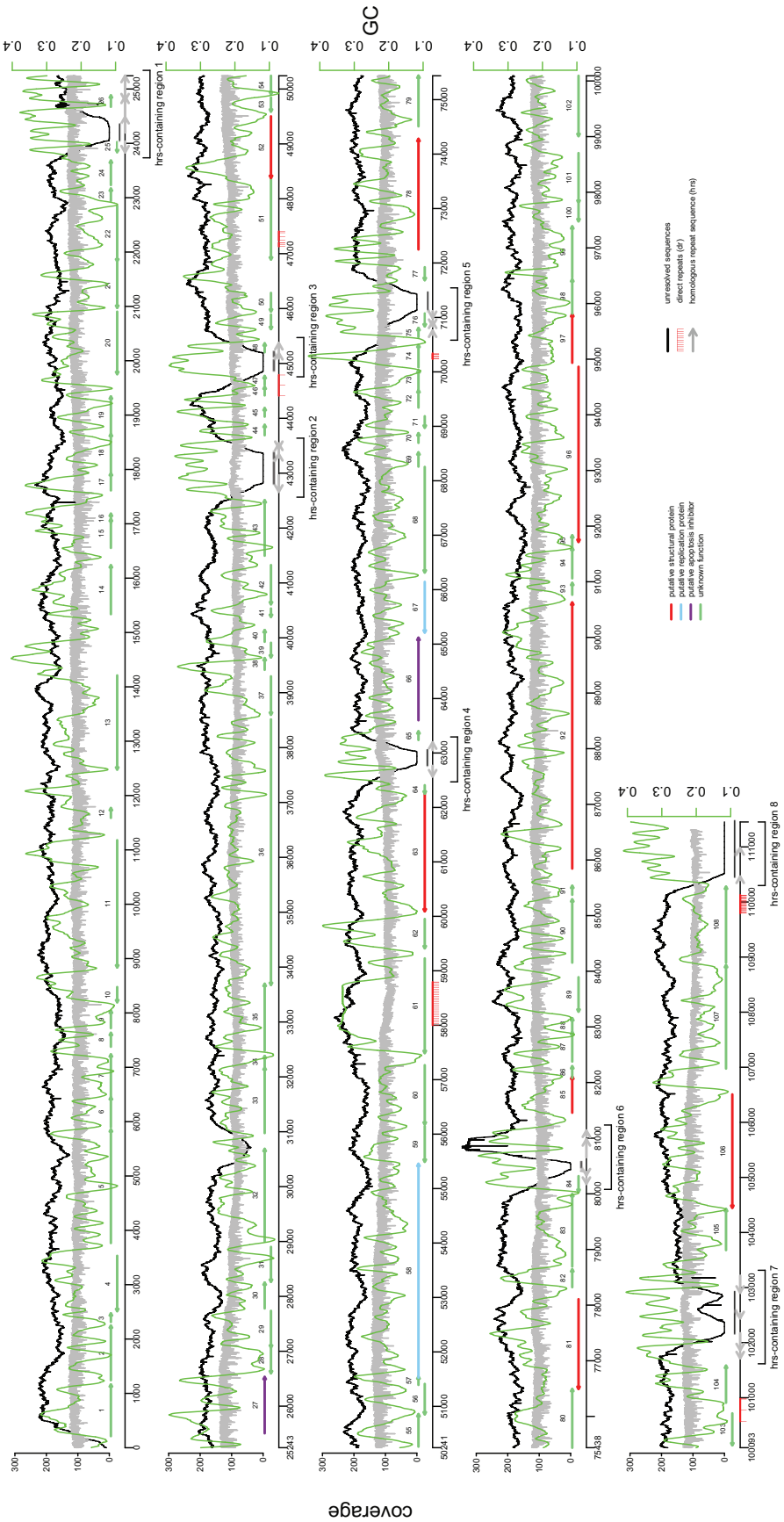

Fig. 6.—Representation of the gene and repeat content of the circular dsDNA genome of LbFV. The black line represents the coverage for Illumina reads and the grey one represents the coverage for the MinION 2D reads (left *y* axis). The green line indicates the GC in a sliding window of 2000 bp (right *y* axis). Unresolved sequences correspond to genomic regions where only MinION reads are available and where polymorphism is very likely.

similar to each other as well as ORF11 and ORF13, suggesting a duplication event took place for these ORFs.

Polymorphism

By mapping the Illumina reads on the final genome sequence, we found clear evidence of single nucleotide polymorphism (SNPs) at two positions, using a minimal allele frequency of 0.1. In position 39,694, the base was either an A (65.4%) or a G (33.3%) out of 162 reads (2 occurrences of T were also observed at that position but this may be the consequence of sequencing error). This A/G polymorphism is expected to translate into amino-acid change in position 73 of ORF39 (Tyr = > Asn). In the intergenic position 85367, we identified an A/T polymorphism (42% A, 58% T out of 176 reads). In addition, apparent mismatches were observed in the reads mapping in the hrs-containing regions and in the repeated region of ORF61. However, it is unclear whether this is a consequence of true polymorphism or to incorrect mapping of the reads in these repeat-containing regions.

Gene Content and Evolutionary History

A total of 108 putative almost nonoverlapping open reading frames (ORFs) starting with a methionine and a minimum length of 50 amino acids were identified (min = 51, max = 1,613, mean = 277.4 amino acids) and distributed equally on both strands (fisher test *P* value = 0.17). All together the predicted ORFs cover 80% (out of 111,453 bp) of the genome which is in the range of what is observed for other arthropods large dsDNA viruses (table 1). Among these ORFs, only 33 have significant protein and/or domains hits (e-value < 0.001) identified by hmmer algorithm (table 2).

Putative Structural Proteins

Based on their hmmer best hits, three LbFV ORFs were identified as putative structural proteins: ORF52 and ORF63 which are homologous to *per os* infectivity factors (Pif) (respectively, pif-2 and p74) and ORF106 which is homologous to Occlusion-Derived Virus (ODV) enveloppe protein. Pif proteins are typically found in arthropod large dsDNA viruses such as Baculoviridae in which they are essential to ensure oral infectivity (Peng et al. 2010; Mu et al. 2014). BLASTp searches (not shown) revealed the presence of a odv-e66 domain in ORF106 (e-value = 6.06e-17).

The closest homologs of these three LbFV ORFs (52, 63 and 106) were found in Hytrosaviridae, although homologs in other arthropod large dsDNA viruses were also detected. Hytrosaviridae is a recently described family of dsDNA viruses, with two representative species infecting the domestic fly *Musca domestica* and the tsetse fly *Glossina pallidipes*. Hytrosaviruses induce a salivary gland hypertrophy symptom although they often persist as a latent asymptomatic infection (Abd-Alla et al. 2008). In Hytrosaviridae, the homologs of the three LbFV proteins are part of the proteome of mature viral particles either in MdSGHV (Garcia-Maruniak et al. 2008) and/or in GpSGHV (Kariithi et al. 2010, 2012) suggesting that ORFs 52, 63 and 106 encode structural proteins of LbFV. ORF106 was predicted to contain a Lyase domain in its N-terminal part. Lyases are enzymes responsible for the breaking of chemical bonds capable of acting on chondroitin, such as Chondroitin sulfate proteoglycans (CSPGs), or hyaluronan. CSPGs are present in various insects' tissues. In the silkworm, where its distribution has been analyzed in detail (Sugiura et al. 2013), CSPGs seem to concentrate at the peritropic membrane where baculovirus infection initiates. Interestingly, it has been found that ODV-e66 has a chondroitinase activity that may facilitate infection during oral infection (Sugiura et al. 2011). The lyase domain identified in ORF106 may thus have a similar chondroitinase property for LbFV.

ORF81 contains an ATPase domain with homologs in archaea, eukaryotes and bacteria and also in Hytrosaviridae. The proteins encoded by the homologs of ORF81 in both species of Hytrosaviruses have been detected in mature particles of MdSGHV and GpSGHV, suggesting that they may also be structural proteins of LbFV.

Based on the hmmer search, ORF78, ORF85, ORF92, ORF96 and ORF97, have either no hits or hits with Hytrosaviridae. However, using a BLASTp search on a custom database containing the arthropod large dsDNA viruses presented in table 1, we found a putative homolog of each of these LbFV proteins in the genomes of both GpSGHV and MdSGHV but not in other virus species. Interestingly, the corresponding hytrosavirus proteins were identified as structural proteins either in GpSGHV or in both MdSGHV and GpSGHV, suggesting that they may represent additional structural proteins of LbFV.

The nine putative structural proteins identified may thus be involved in the high infectivity of LbFV under conditions of superparasitism. They may also be involved in vertical transmission efficiency since vertical transmission may imply the recolonization of the developing embryo by LbFV particles that are simultaneously injected into the *Drosophila* larva.

ORF61 Has a Mucin-like Structure

As previously mentioned, ORF61 is a high molecular weight protein containing 23.5 repetitions of a threonine rich motif (TTTTTSTTLKP) which makes it a mucin-like protein. Mucins have typically high molecular weight and contain a large central region formed of multiple tandem repeats of a serine or threonine-rich motif which is usually heavily glycosylated. They are found in eukaryotes but also in some viruses, such as in Filoviridae (Hashiguchi et al. 2015), Herpesviridae (Altgärde et al. 2015), Paramyxoviridae (respiratory syncitial virus, RSV), and HIV, where they are incorporated into the envelope. ORF61 may thus also be a structural protein of LbFV. Mucin-like proteins found in viruses mediate the binding with host cell surface by interacting with host

Table 2

ORF prediction in LbFV genome

ORF #	O	Start	Stop	Length (aa)	Hits found	Top Hit accession number	Description	e-value	Top hit species	Kingdom	Accession	Clan	Description	Start	End	Ind-e-value	Cond-e-value	TM	SP	struct. MdSGHV	struct. GpSGHV
ORF1	+	214	1164	317	585	W6JLD1_9POXV	Translation elongation factor eEF-3 like	3.50E-34	Anomala cuprea entomopoxvirus	Virus	PF14388.4	n/a	Domain of unknown function (DUF4419)	26	311	6.60E-45	4.00E-49				
ORF2	+	1243	2247	335	0																
ORF3	+	2314	2472	53	0																
ORF4	–	2514	3530	339	0																
ORF5	+	3767	5857	697	1	A0A084WAW3_ANOSI	Nav1.4a	5.70E-06	Culex quinquefasciatus	Eukaryote								2			
ORF6	+	5899	6450	184	0																
ORF7	+	6455	7237	261	0																
ORF8	+	7379	7633	85	0																
ORF9	+	7721	8041	107	0																
ORF10	–	8197	8475	93	2304	C8AXE7_9CICH	TMO-4c4 (Fragment)	1.30E-23	Oreochromis niloticus	Eukaryote											
ORF11	–	8838	11177	780	94	B4MZD3_DROWI	Uncharacterized protein	2.70E-17	Lucilia cuprina	Eukaryote	PF02373.20	CL0029	JmjC domain, hydroxylase	475	602	5.00E-09	3.10E-13				
ORF12	+	11590	11769	60	0																
ORF13	–	12479	14209	577	929	A0A0P6K4V2_HETGA	V-type proton ATPase subunit a	2.10E-20	Fukomys damarensis	Eukaryote	PF02373.20	CL0029	JmjC domain, hydroxylase	283	403	1.20E-10	7.40E-15				
ORF14	+	15338	16237	300	169	A0A0B5A143_9ABAC	Uncharacterized protein	1.10E-21	Pseudoplusia includens SNPV IE	Virus	PF02498.15	n/a	BRO family, N-terminal domain	97	195	5.50E-14	3.40E-18				
ORF15	+	16558	17091	178	0																
ORF16	+	17012	17185	58	0																
ORF17	+	17615	17932	106	0																
ORF18	+	17954	18655	234	0																
ORF19	+	18662	19342	227	0																
ORF20	–	19754	20902	383	0																
ORF21	–	20974	21804	277	0																
ORF22	–	21829	22842	338	0																
ORF23	+	22941	23168	76	0																
ORF24	+	23234	23683	150	0																
ORF25	–	23838	24026	63	0																
ORF26	+	24673	24882	70	0																
ORF27	+	25506	26531	342	3883	A0A182XPW4_ANOQN	Uncharacterized protein	5.00E-100	Anopheles christyi	Eukaryote	PF00653.19	CL0417	Inhibitor of Apoptosis domain	11	78	3.50E-15	4.30E-19				
											PF00653.19	CL0417	Inhibitor of Apoptosis domain	172	237	4.90E-16	6.00E-20				
											PF13920.4	CL0229	Zinc finger, C3HC4 type (RING finger)	290	334	1.10E-11	1.30E-15				

(continued)

Table 2 Continued

ORF #	O	Start	Stop	Length (aa)	Hits found	Top Hit accession number	Description	e-value	Top hit species	Kingdom	Accession	Clan	Description	Start	End	Ind-e-value	Cond-e-value	TM	SP	struct. MdSGHV	struct. GpSGHV
ORF28	–	26599	27075	159	0																
ORF29	–	27068	27736	223	0																
ORF30	+	27781	28251	157	0																
ORF31	–	28266	28901	212	0																
ORF32	+	29002	30687	562	0																
ORF33	+	30976	32172	399	16	Q9YVQ5_MSEPV	ORF MSV187 putative late transcription factor VLTF-2 homolog (Vaccinia A1L), similar to SW:P33814	9.20E-13	Melanoplus sanguinipes entomopoxvirus	Virus											
ORF34	+	32165	32380	72	0																
ORF35	+	32481	33695	405	7	Q9YVQ5_MSEPV	ORF MSV187 putative late transcription factor VLTF-2 homolog (Vaccinia A1L), similar to SW:P33814	6.50E-09	Melanoplus sanguinipes entomopoxvirus	Virus											
ORF36	–	33675	38513	1613	0																
ORF37	–	38599	39297	233	106	Q7T5L7_GVCL	Dbp	3.20E-15	Cryptophlebia leucotreta granulosis virus	Virus	PF00293.26	CL0261	NUDIX domain	23	165	6.20E-09	3.80E-13				
ORF38	+	39418	39630	71	0																
ORF39	–	39643	39909	89	0																
ORF40	+	39929	40141	71	0																
ORF41	–	40381	40533	51	0																
ORF42	–	40611	41324	238	0																
ORF43	+	41497	42507	337	0																
ORF44	+	43692	43880	63	0																
ORF45	+	44031	44186	52	1	A6YEY0_9NEOP	NADH-ubiquinone oxidoreductase chain 1 (Fragment)	0.00018	Operophtera brumata	Eukaryote											
ORF46	+	44418	44576	53	0																
ORF47	+	44612	44770	53	0																
ORF48	+	45204	45365	54	0																
ORF49	–	45626	45889	88	166	W4XJ30_STRPU	Uncharacterized protein	2.20E-06	Strongylocentrotus purpuratus	Eukaryote	PF01712.17	CL0023	Deoxynucleoside kinase	1	86	1.90E-11	1.20E-15				
ORF50	–	45946	46284	113	120	D9YXA4_9CRUS	NADH-ubiquinone oxidoreductase chain 2 (Fragment)	4.60E-10	Daphnia magna	Eukaryote	PF01712.17	CL0023	Deoxynucleoside kinase	5	111	2.70E-16	1.70E-20				
ORF51	–	46894	48393	500	0																

(continued)

Table 2 Continued

ORF #	O	Start	Stop	Length (aa)	Hits found	Top Hit accession number	Description	e-value	Top hit species	Kingdom	Accession	Clan	Description	Start	End	Ind-e-value	Cond-e-value	TM	SP	struct. MdSGHV	struct. GpSGHV
ORF52	–	48380	49501	374	37	A0A120HYA7_GHVS	Per-os infectivity factor 2-like protein	5.30E-13	Glossina hytrovirus	Virus								1			x
ORF53	–	49589	49843	85	0																
ORF54	–	49905	50240	112	2	A0A0A1HB82_LINLA	Cytochrome c oxidase subunit 1 (Fragment)	6.00E-13	Strongylocentrotus purpuratus	Eukaryote								2			
ORF55	+	50261	50857	199	0																
ORF56	–	50854	51411	186	0																
ORF57	+	51395	51547	51	0																
ORF58	+	51528	55454	1309	3	A0A0K1L634_9VIRU	ORF_007L	6.30E-05	Scale drop disease virus Iridoviridae	Virus	PF00136.19	CL0194	DNA polymerase family B	599	880	2.30E-20	1.40E-24	2			
ORF59	–	55493	56143	217	0																
ORF60	–	56177	57265	363	4	R9XIV9_ASHAC	AaceriAGL139Wp	4.70E-09	Glossina hytrovirus	Virus	PF02450.13	CL0028	Lecithin:cholesterol acyltransferase	57	351	1.60E-07	9.60E-12				
ORF61	–	57490	59226	579	476	B4JWM1_DROGR	GH22716	4.80E-78	Drosophila melanogaster	Eukaryote			mucin						1		
ORF62	–	59414	59950	179	0																
ORF63	–	60093	62282	730	153	B0YLJ3_GHVS	p74 protein-like protein	1.30E-68	Glossina hytrovirus	Virus										x	x
ORF64	–	62255	62407	51	0																
ORF65	+	63240	63398	53	0																
ORF66	+	63602	65116	505	1892	B4NG60_DROWI	Uncharacterized protein	1.30E-50	Lucilia cuprina	Eukaryote	PF00653.19	CL0417	Inhibitor of Apoptosis domain	123	189	2.40E-12	2.90E-16				
											PF00653.19	CL0417	Inhibitor of Apoptosis domain	290	355	1.60E-14	2.00E-18				
											PF13920.4	CL0229	Zinc finger, C3HC4 type (RING finger)	396	442	2.30E-08	2.80E-12				
											PF13920.4	CL0229	Zinc finger, C3HC4 type (RING finger)	452	495	1.60E-06	2.00E-10				
ORF67	–	65213	66142	310	4	A0A0U4AXZ3_9VIRU	Putative helicase	2.30E-06	Pithovirus sibericum	Virus											
ORF68	–	66316	68253	646	0																
ORF69	+	68261	68521	87	0																
ORF70	+	68697	68873	59	0																
ORF71	–	68975	69184	70	0																
ORF72	+	69345	69665	107	0																
ORF73	+	69662	70030	123	0																
ORF74	+	70085	70495	137	0																
ORF75	+	70641	70805	55	0																

(continued)

Table 2 Continued

ORF #	O	Start	Stop	Length (aa)	Hits found	Top Hit accession number	Description	e-value	Top hit species	Kingdom	Accession	Clan	Description	Start	End	Ind-e-value	Cond-e-value	TM	SP	struct. MdSGHV	struct. GpSGHV
ORF76	–	70836	71072	79	0																
ORF77	–	71685	71921	79	0																
ORF78	+	72249	74279	677	5	A0A0R3WJU6_HYDTA	Uncharacterized protein	1.60E-09	Glossina hytrovirus	Virus								1			x
ORF79	+	74511	75440	310	0																
ORF80	+	75437	76492	352	0																
ORF81	–	76495	78105	537	5912	B0YLU3_GHVS	Uncharacterized protein	2.20E-18	Glossina hytrovirus	Virus	PF00004.27	CL0023	ATPase associated with various activities (AAA)	202	331	4.40E-14	2.70E-18			x	x
ORF82	+	78323	78643	107	0																
ORF83	+	78694	79995	434	0																
ORF84	–	80036	80317	94	0																
ORF85	+	81459	82106	216	77	B2YG85_MHVB	Ac81-like protein	6.70E-10	Musca hytrovirus	Virus	PF05820.9	n/a	Baculovirus protein of unknown function (DUF845)	29	147	0.00015	9.10E-09	2			x
ORF86	+	82084	82293	70	0																
ORF87	+	82378	82908	177	0																
ORF88	+	82908	83156	83	0																
ORF89	-	83275	83892	206	2	B6DZA6_9VIRU	Uncharacterized protein	4.40E-107	Leptopilina boulardi filamentous virus	Virus									1		
ORF90	+	84156	85292	379	0																
ORF91	+	85375	85536	54	0																
ORF92	+	85846	90627	1594	2	B0YLK7_GHVS	Per-os infectivity factor 2-like protein	3.30E-18	Glossina hytrovirus	Virus								4			x
ORF93	+	90767	90979	71	0																
ORF94	+	91062	91610	183	0																
ORF95	+	91631	91834	68	1	A0A061DCU0_BABBI	Uncharacterized protein	0.00012	Theileria annulata	Eukaryote								1			
ORF96	–	91717	94863	1049	2	B0YLT7_GHVS	Putative uncharacterized protein	5.50E-28	Glossina hytrovirus	Virus											x
ORF97	+	94936	95808	291	0															x	x
ORF98	+	95820	96437	206	0																
ORF99	+	96467	97387	307	0																
ORF100	–	97479	97787	103	0																

(continued)

Table 2 Continued

ORF #	O	Start	Stop	Length (aa)	Hits found	Top Hit accession number	Description	e-value	Top hit species	Kingdom	Accession	Clan	Description	Start	End	Ind-e-value	Cond-e-value	TM	SP	struct. MdSGHV	struct. GpSGHV
ORF101	−	97813	98703	297	1	R9X889_ASHAC	Vacuolar ATPase assembly integral membrane protein VMA21	6.70E-06	Musca hytrovirus	Virus	PF10553.7	n/a	MSV199 domain	23	101	5.40E-05	3.30E-09	1			
ORF102	−	99010	100092	361	0																
ORF103	−	100137	100715	193	0																
ORF104	+	100910	101581	224	0																
ORF105	+	103681	104427	249	1	A0A016TCB2_9BILA	Hexosyltransferase	2.40E-06	Caenorhabditis brenneri	Eukaryote											
ORF106	−	104449	106506	686	138	A0A0B7JV03_BIOOC	Occlusion-derived virus envelope protein	1.10E-50	Musca hytrovirus	Virus	PF08124.9	CL0372	Polysaccharide lyase family 8, N terminal alpha-helical domain	56	283	1.30E-10	8.20E-15	2		x	
ORF107	+	106977	108854	626	1	A0A158R6B8_9BILA	Uncharacterized protein	0.00023	Helobdella robusta	Eukaryote											
ORF108	+	108899	110278	460	28	G5AXB1_HETGA	Protein CASP	2.10E-22	Fukomys damarensis	Eukaryote											

Note.—The annotation was done using hmmer algorithm for protein (phmmer) and domain (hmmscan) searches.

O: Orientation; SP, signal peptide; TM, transmembrane domain; Struct. MdSGHV: homolog has been identified as a structural protein in MdSGHV; Struct. GpSGHV: homolog has been identified as a structural protein in GpSGHV.

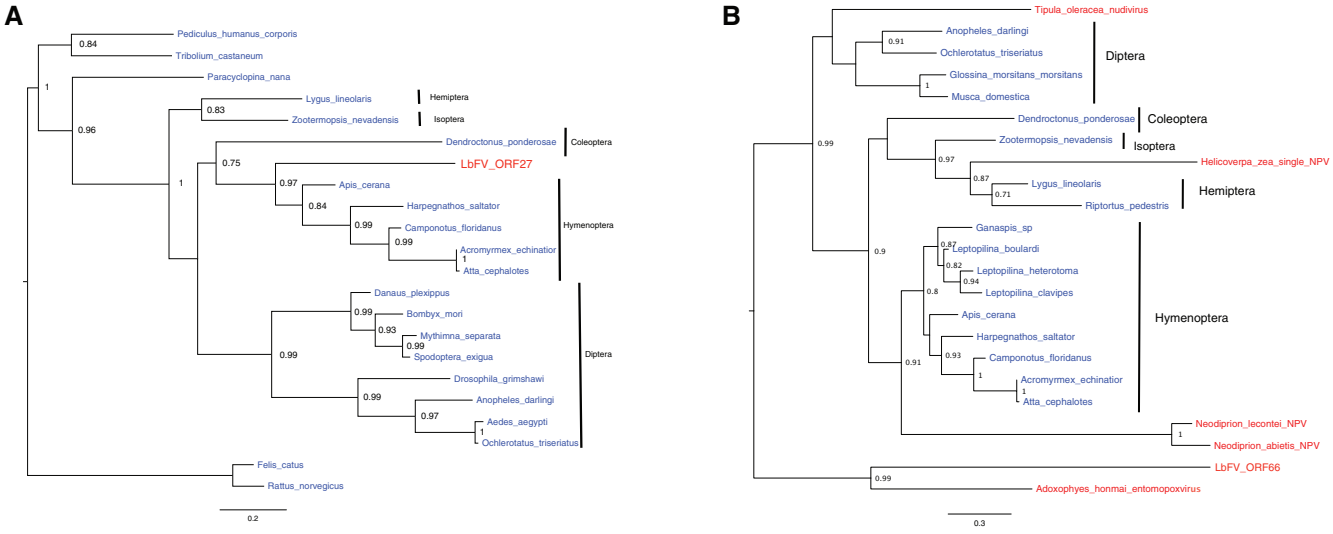

Fig. 7.—The phylogeny of two putative anti-apoptotic genes found in LbFV genome suggest that they have been acquired from eukaryotes. (*A*) ORF27 and (*B*) ORF66. The branch supports were estimated by approximate Likelihood ratio tests (aLRT). Only aLRT > 0.70 are represented. Eukaryotic lineages are represented in blue and virus lineages are in red.

glycosaminoglycans, such as chondroitin sulfate and sulfated hyaluronan (Altgärde et al. 2015). Because ORF61 has a putative glycosaminoglycan-ligand activity and ORF106 has putative glycosaminoglycan-breaking activity, we can speculate that both ORFs may work in conjunction to facilitate LbFV entry into the cell. Alternatively or in addition, the glycosaminoglycan-breaking activity of ORF106 may be required for releasing the virus from its interaction with host cell surface glycosaminoglycans induced by ORF61 to ensure particle release. ORF61 contains a signal peptide on its N-terminal region (positions 1–24).

A DNA Polymerase but No RNA Polymerase Identified

A putative DNA polymerase (ORF58) related to DNApol B found in some archaea, eukaryotes and large dsDNAviruses was identified (table 2) together with a putative helicase (ORF67).

Like those of most large DNA viruses, the LbFV genome was expected to encode a DNA-dependent RNA polymerase for the transcription of its genes. This function is normally fulfilled by a complex of proteins including at least two RNA polymerase subunits and initiation, elongation, and termination factors. Like in Hytrosaviridae sequenced so far, none of the 108 LbFV ORFs showed similarity with known DNA-dependent RNA polymerase subunits. Either the host RNA polymerase is responsible for all the transcription LbFV genes, or a yet unidentified complex of proteins encoded by the genome of LbFV fulfills this function.

bro Proteins

ORF14 showed similarity with bro proteins which are typically found in baculoviruses and other large dsDNA viruses, as well as in dsDNA phages and prokaryotic transposons (Gauthier et al. 2015). The function of bro proteins is still unknown, although it has been suggested that BRO-A and BRO-C are DNA binding proteins that influence host DNA replication and/or transcription in baculoviruses (Zemskov et al. 2000). Although normally found in several copies in the genomes, only one representant of bro protein was identified in the genome of LbFV.

Two Inhibitors of Apoptosis Were Acquired from Eukaryotes

As often found in large dsDNA viruses, we found two putative ORFs (ORF27, ORF66) with similarities with inhibitors of apoptosis. The phylogenetic analysis of both genes suggests an horizontal acquisition from eukaryotes (fig. 7; supplementary table S4, Supplementary Material online). For ORF27, all hymenoptera sequences clusters within a monophyletic clade that includes ORF27 in a basal position. This suggests that LbFV acquired this anti-apoptotic gene by horizontal transfer from an insect host, most likely an Hymenoptera (fig. 7*A*). The second putative inhibitor of apoptosis encoded in the genome of LbFV (ORF66) clusters with an entomopoxvirus (fig. 7*B*). All other sequences ($n = 17$) that we were able to include in the analysis, except 4, are insect sequences, thus suggesting an horizontal acquisition from an unidentified ancestral insect host. Interestingly, the phylogeny suggests other cases of horizontal acquisition from insect hosts by large dsDNA viruses. *Tipula oleracea* nudivirus which is known to infect diptera seems to have acquired its inhibitor of apoptosis gene from a diptera, as could be expected. *Neodiprion* NPVs which infect Hymenoptera form a monophyletic clade relatively well supported (aLRT = 0.91) with Hymenopteras, suggesting that it

acquired its anti-apoptotic gene from an ancestral Hymenoptera. Finally, *Helicoverpa zea* sNPV forms a cluster relatively well supported (0.87) with two Hemipteran species which may suggest that HzNPV acquired this gene from an Hemipteran. This may sound surprising since nowadays *Helicoverpa zea* sNPV infects Lepidoptera, even if host switches may have occurred in the past explaining the pattern. The repeated acquisition of inhibitors of apoptosis by dsDNA viruses most likely reflects an adaptation of these intra-cellular parasites to manipulating the host immune defense that consists in eliminating infected cells.

ORF60 May Be Involved in Membrane Fusion

ORF60 contains a putative Lecithin cholesterol acyltransferase domain. This domain is shared by proteins found in numerous prokaryotes eukaryotes, archaea and in some viruses. To date it has been identified only, to our knowledge, in Hytrosaviridae in both representants of this group MdSGHV (Garcia-Maruniak et al. 2008) and GpSGHV (Kariithi et al. 2010). Lecithin cholesterol acyltransferase (LCAT) is involved in high density lipoprotein (HDL) maturation in mammals (Piper et al. 2015) and may be involved in membrane fusion in viruses.

LbFV Captured a Gene with Jumonji C Domain from Its Host

Interestingly, two genes (ORF11 and ORF13) share a Jumonji C (JmJC) domain. JmjC containing enzymes represent the most important class of demethylase enzymes and catalyse lysine demethylation of histones (Klose et al. 2006). They are thus involved in the regulation of the transcription. Different classes of JmJC proteins have been documented in eukaryotes and prokaryotes. In addition, it has been reported that a couple of viruses with very large genome size (pandoravirus and megavirus) encode proteins containing JmJC domain. However, there was no detectable similarity between these viral JmJC–containing proteins and the two proteins found in LbFV (ORF11 and ORF13, not shown). Instead we found only eukaryotic homologs in the databases (including *Leptopilina* species) that we aligned and used to construct a phylogeny (fig. 8 and supplementary table S4, Supplementary Material online). Interestingly, the phylogeny suggests that LbFV captured an ancestral JmJC-containing gene from its insect host and that afterwards a duplication event took place leading to two divergent copies of the JmJC-containing gene in the genome of LbFV. The horizontal transfer probably involved an ancestral Hymenoptera since the branch leading to the two LbFV ORFs is nested within a well supported clade of Hymenoptera (aLRT = 0.92).

Other Putative Functions

Other putative, functions were identified: phosphohydrolase (ORF37), kinase (ORF49 and ORF50), ATPase (ORF81). Finally ORF101 contains a MSV199 domain which may have DNA binding properties (Iyer et al. 2002). ORF85 showed similarity with the baculovirus core gene of the protein Ac81 which function is unknown. Finally, there was no evidence of the presence of tRNAs in the genome, according to the softwares t-RNAscan-SE (Lowe and Eddy 1997) and ARAGORN (Laslett & Canback 2004).

Promoter Analysis

Because promoter analysis may help identifying groups of genes with similar transcription regulation, we searched for conserved motifs in the 300bp upstream of each ORF. Six motifs with variable levels of conservation (mean identity between 51 and 98%) were identified by MEME (supplementary fig. S4 and table S3, Supplementary Material online). In particular, we found a 11bp long TATA repetition (motif #4) in 32 ORFs, sometimes in several copies. No conserved motif was found for 28 ORFs. None of the six identified motifs were particularly associated with the nine genes encoding the putative structural proteins (glm models, all P value >0.05).

Phylogenetic Position of LbFV

We first searched for homologs to the 108 LbFV ORFs in the generalist database Uniprot using hmmer algorithm. Among the 33 for which we obtained a hit, only 18 had their best hit with known viruses. Among these 18 ORFs, ten of them had their best hit with either Musca domestica SGHV or Glossina pallidipes SGHV. The other viral best-hmmer hits point towards other large dsDNA viruses (baculoviruses: ORF14, ORF37, entomopoxviruses: ORF1, ORF33, ORF35, pithovirus ORF67, iridoviridae ORF58).

To ascertain the position of LbFV within the phylogeny of arthropods large dsDNA viruses, we searched for LbFV orthologs for each of the 108 LbFV predicted proteins within the predicted proteome of 13 representative arthropods dsDNA viruses (table 1). Most of the best hits (17/32) were obtained with Hytrosaviridae. Six LbFV genes had at least four orthologs in four virus species from our custom database. These six genes were used to reconstruct the phylogeny of the 14 viruses (supplementary table S5, Supplementary Material online). Three of these genes were putative structural genes (ORF52, ORF63 and ORF106), one is a homolog of Ac81 found in all baculovirus genomes, nudivirus and Hytrosaviridae (ORF85) (Rohrmann 2014), one is a putative DNA polymerase (ORF58) and one is a putative phosphohydrolase (ORF37). Based on the concatenated alignment of conserved blocks identified in the six ORFs, we constructed a phylogeny based on maximum likelihood. The phylogeny obtained (fig. 9) was globally well resolved and each family Hytrosaviridae, Nudiviridae and Baculoviridae were found to be monophyletic with the expected relationship among them (Bézier et al. 2015). From this analysis, the position of the recently sequenced Filamentous virus of *Apis mellifera* was unclear, but

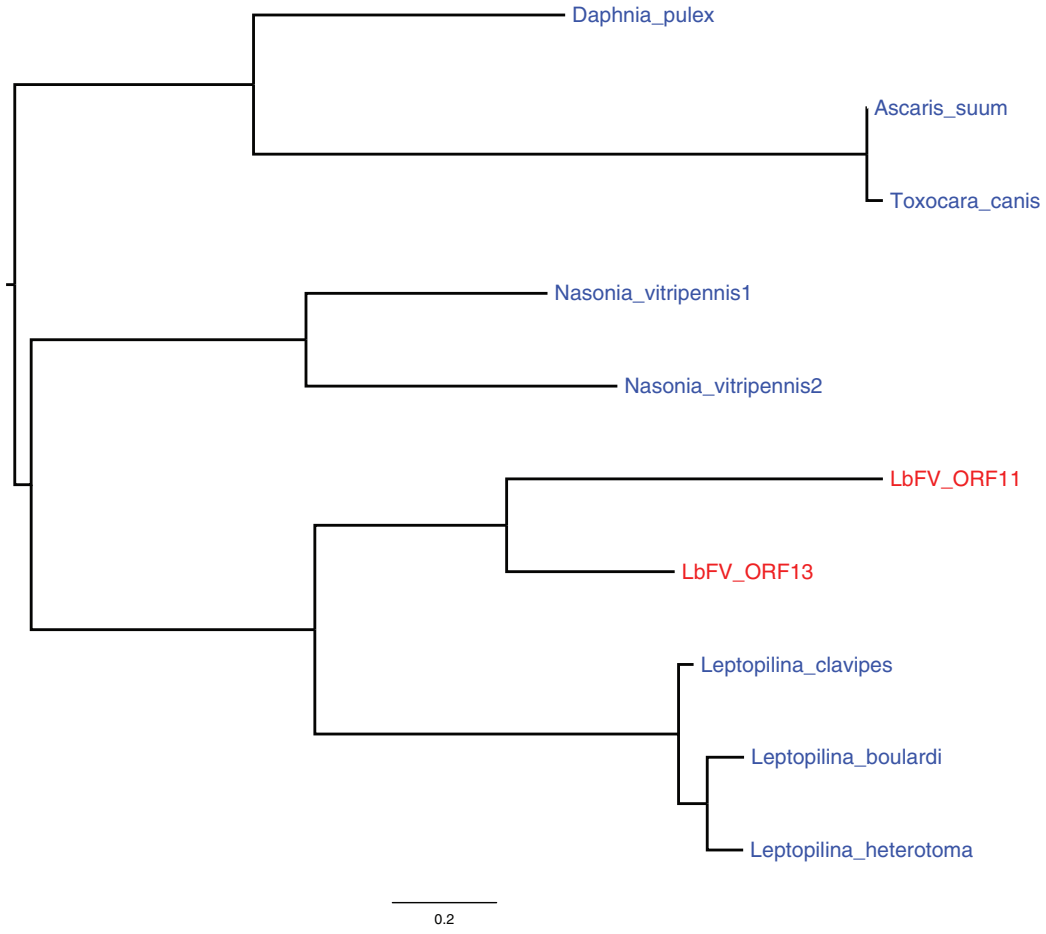

Fig. 8.—The phylogeny of LbFV ORFs 11 and 13 suggests that they have been horizontally acquired from eukaryotes and evolved subsequently through a duplication in the genome of LbFV. The branch supports were estimated by approximate Likelihood ratio tests (aLRT). Only aLRT > 0.70 are represented. Eukaryotic lineages are represented in blue and virus lineages are in red.

appeared to be very distantly related to LbFV. LbFV clustered within a very well supported clade (aLRT = 1) including the two Hytrosaviridae genomes sequenced so far. Altogether the results suggest that LbFV is related to Hytrosaviridae. However, LbFV is phylogenetically distant from them and may thus be the representant of a new virus family.

Discussion–Conclusion

We presented here the genome of the behavior manipulating virus LbFV. This DNA virus is characterized by a very high AT content (~80%), encodes for 108 putative ORFs, with few (33/108) having homologs in public database. The genome is ~111 kb long, circular and contains an homologous repeated sequence (*hrs*) located in eight genomic regions of the genome sometimes in several copies. The presence of homologous segments has been reported in other viruses and is indeed considered to increase genome plasticity, facilitating the intra and intermolecular recombination (Imperiale & Jiang 2015). In particular, the presence of two inverted repeats flanking segments of unique sequences is responsible for frequent inversions in the genomes of Herpesviridae. As a result, the herpesvirus population consists of an equimolar mixture of four virus recombinants (Bataille & Epstein 1995; McVoy & Ramnarain 2000). In this work, we found some evidence of recombination among the hrs regions. These evidences come from mapping the 14 millions Miseq reads on the contigs and from PCR assays. The presence of pairs of reads mapping to alternative contigs suggests that there is some recombinant viruses in the population. However, those alternative variants may be in extremely low frequency in the population, explaining why only 96 Illumina reads were detected (out of 14 million reads mapped). The PCR assay confirms the presence of multiple connexions among the eight identified blocks but again does not inform on the frequency of such recombinants. Surprisingly, the addition of long reads (MinION) revealed that the LbFV population is in fact composed of one unique major genomic form with a unique organization of the eight blocks on the chromosome.

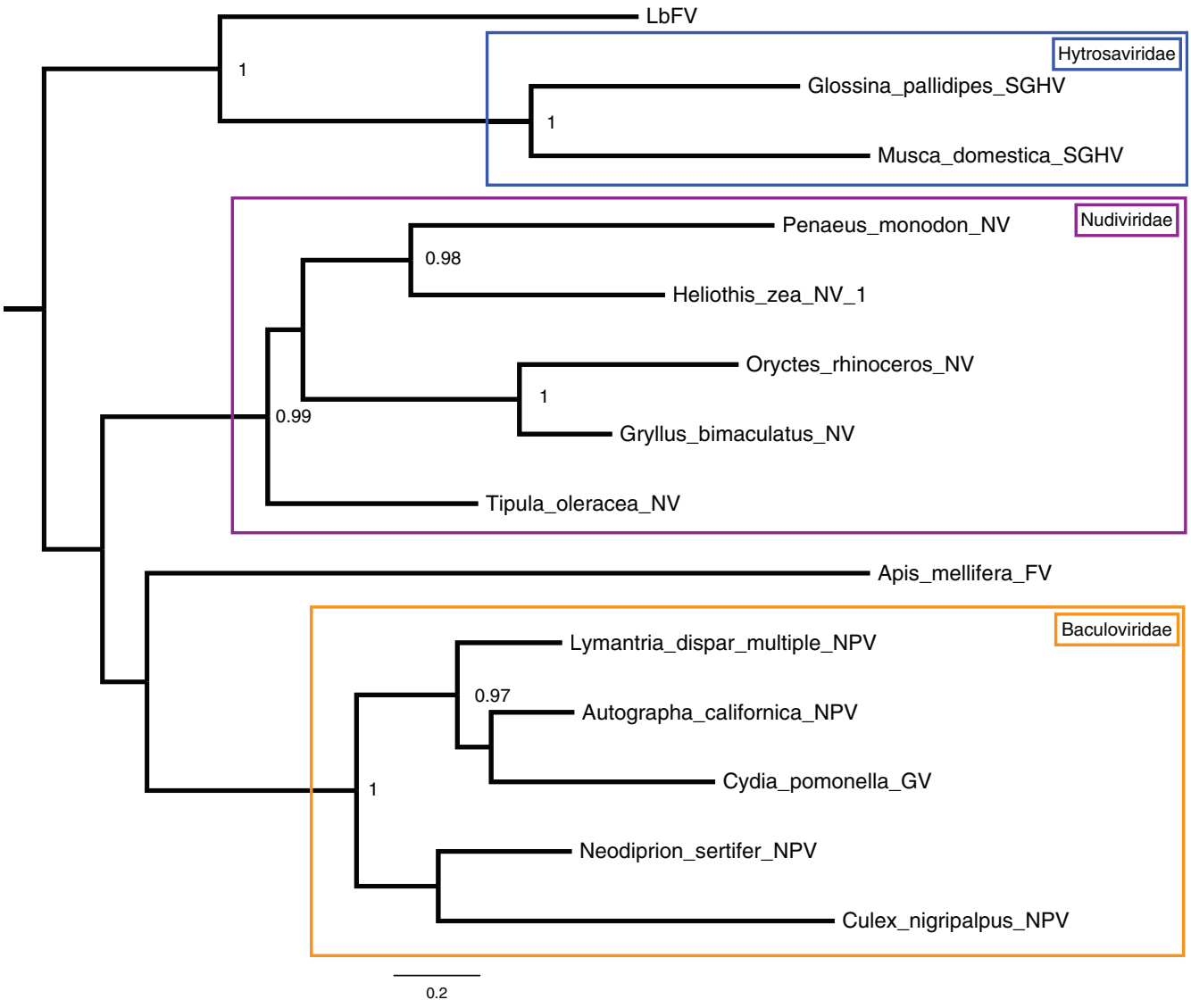

Fig. 9.—Phylogenetic reconstruction based on the concatenation of 6 genes (ORF37, ORF52, ORF58, ORF85 and ORF106). The conserved blocks were concatenated and the phylogeny was built using PhyML. The phylogenetic signal was congruent among the six genes. The branch supports were estimated by approximate Likelihood ratio tests (aLRT). Only aLRT > 0.70 are represented.

However, because the output of MinION is orders of magnitude below that of Illumina sequencing (~8,000 MinION reads vs. 14 million Illumina reads), we cannot exclude that there are some genomic rearrangements segregating at low frequency in the population. Importantly, the PCR assay also indicated that there is polymorphism in the genomic regions separating the eight blocks since several bands were obtained for some connections (fig. 5C). Note that for some other connexions we did not get any PCR product which may be explained by the highly stable secondary structures that is predicted in these hrs-containing regions (supplementary fig. S5, Supplementary Material online). The coexistence of several variants even within single individuals is well known for some large dsDNA viruses, such as the white spot syndrome virus (WSSV) in regions with variable number tandem repeats (VNTR, Hoa et al. 2011). Because these regions in LbFV genome contain repetitions of the hrs sequence we have to rely on the long reads to decipher them since Illumina reads are too short to be correctly mapped in these regions. However, since MinION reads have a high error rate, we cannot either confirm or infirm the presence of polymorphism in these regions because the variation observed may be due to error rate and/or to true polymorphism. However, these data suggest that the virus population is composed of a major genomic organization with low-frequency rearranged genomes. In addition, the presence of polymorphism in the unresolved

hrs-containing regions is very likely, although again the frequency of the minor alleles is unknown. This raises the question of the adaptive significance of this polymorphism and we may speculate that it could allow the virus to quickly adapt to new environments, such as new parasitoid lineages.

The phylogenetic analysis clearly indicates that LbFV is related to Hytrosaviridae with which it forms a monophyletic group. However, LbFV is only distantly related to Hytrosaviridae and may thus represent a new virus family. Genomic data on additional related viruses will be necessary to answer this question. The evolutionary origin of LbFV is unclear but it is interesting to note that the related Hytrosaviridae infect Diptera. Whether the ancestor of LbFV infected a Diptera, like *Drosophila*, is obviously a tempting hypothesis since host-parasitoid relationship may favor the occurrence of horizontal transfer (Renault et al. 2005). Alternatively, we can imagine that LbFV is a representant of a new family of insect viruses that ancestrally infected Hymenoptera.

The question of how parasites manages to manipulate the behavior of their hosts is an open intriguing question (Van Houte et al. 2013). To our knowledge, parasites genes directly or indirectly involved in the behavioral manipulation have been unequivoquely identified only in a few systems involving baculoviruses responsible for the tree-top disease of their caterpillar hosts: when the caterpillar is infected, it climbs to the top of the plant where it eventually dies and liquefies, thus releasing its viral particles. This behavior modification is supposed to enhance baculovirus dispersal. Interestingly, one of the gene identified (ptp gene) as being involved in the manipulation has been acquired through horizontal transfer from the Lepidopteran host (Hoover et al. 2011; Katsuma et al. 2012). Similarly, we found that the genome of LbFV encodes two proteins with a predicted JmJC domain (ORF 11 and 13) and that this gene has been captured from a wasp ancestor followed by gene duplication (fig. 8). Proteins containing JmjC domain are predicted to be metalloenzymes that adopt the cupin fold and are candidates for enzymes that regulate chromatin remodeling. JmjC domains have been identified in numerous eukaryotic proteins containing domains typical of transcription factors and we may speculate that this gene mediates the manipulation of wasp gene expression, possibly in relation with the behavioral manipulation. Interestingly, we found that ORF13 is upregulated in the head of the wasp compared with the expression in the abdomen (Varaldi J et al. unpublished). All these elements make ORF13 a good candidate for being a gene that have been co-opted by the virus to manipulate the behavior of the wasp. Additional functional tests of this hypothesis are obviously required.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Catherine Hänni for the access to the GS junior and ProfileExpert for the MiSeq sequencing. The MinION sequencing was performed at the sequencing platform of the IGFL. We thank the DROPARCON consortium for sharing sequence data and Van Tran-Van for help in the PFGE. This work was supported by the Centre National de la Recherche Scientifique (UMR CNRS 5558) and by the Agence Nationale de la Recherche (ANR) (11-JSV7-0011 Viromics). The bioinformatic work was performed using the computing facilities of the CC LBBE/PRABI.

Literature Cited

Abd-Alla AMM, et al. 2008. Genome analysis of a *Glossina pallidipes* salivary gland hypertrophy virus reveals a novel, large, double-stranded circular DNA virus. J Virol. 82:4595–4611.

Abd-Alla AMM, et al. 2009. Hytrosaviridae: a proposal for classification and nomenclature of a new insect virus family. Arch Virol. 154:909–918.

Altgärde N, et al. 2015. Mucin-like region of herpes simplex virus type 1 attachment protein glycoprotein C (gC) modulates the virus-glycosaminoglycan interaction. J Biol Chem. 290:21473–21485.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:W202–W208.

Bataille D, Epstein AL. 1995. Herpes simplex virus type 1 replication and recombination. Biochimie. 77:787–795.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573.

Bézier A, et al. 2015. The genome of the nucleopolyhedrosis-causing virus from *Tipula oleracea* sheds new light on the Nudiviridae family McFadden, G, editor. J Virol. 89:3008–3025.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Biological and medical physics, biomedical engineering. Bastolla U, Porto M, Roman HE and Vendruscolo M, editors. Springer Berlin Heidelberg: Berlin, Heidelberg. p. 207–232.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–W37.

Gandon S, Rivero A, Varaldi J. 2006. Superparasitism evolution: adaptation or manipulation? Am Nat. 167:E1–E22.

Garcia-Maruniak A, Maruniak JE, Farmerie W, Boucias DG. 2008. Sequence analysis of a non-classified, non-occluded DNA virus that causes salivary gland hypertrophy of Musca domestica, MdSGHV. Virology 377:184–196.

Gauthier L, et al. 2015. The *Apis mellifera* filamentous virus genome. Viruses 7:3798–3815.

Godfray H. 1994. Parasitoids: behavioral and evolutionary ecology. Princeton, New Jersey: Princeton University Press.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27:221–224.

Hashiguchi T, et al. 2015. Structural basis for Marburg virus neutralization by a cross-reactive human antibody. Cell 160:904–912.

Hilton S, Winstanley D. 2008. The origins of replication of granuloviruses. Arch Virol. 153:1527–1535.
Hoa TTT, et al. 2011. Mixed-genotype white spot syndrome virus infections of shrimp are inversely correlated with disease outbreaks in ponds. J Gen Virol. 92:675–680.
Hoover K, et al. 2011. A gene for an extended phenotype. Science 333:1401.
Hunt M, et al. 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14:R47.
Imperiale MJ, Jiang M. 2015. What DNA viral genomic rearrangements tell us about persistence. J Virol. 89:1948–1950.
Iyer LM, Koonin EV, Aravind L. 2002. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol. 3:research0012.1.
Jain M, et al. 2015. Improved data analysis for the MinION nanopore sequencer. Nat Methods 12:351–356.
Kariithi HM, et al. 2010. Proteomic analysis of Glossina pallidipes salivary gland hypertrophy virus virions for immune intervention in tsetse fly colonies. J Gen Virol. 91:3065–3074.
Kariithi HM, et al. 2012. Correlation between structure, protein composition, morphogenesis and cytopathology of *Glossina pallidipes* salivary gland hypertrophy virus. J Gen Virol. 94:193–208.
Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Katsuma S, et al. 2012. Baculovirus-encoded protein BV/ODV-E26 determines tissue tropism and virulence in lepidopteran insects. J Virol. 86:2545–2555.
Klose RJ, Kallin EM, Zhang Y. 2006. JmjC-domain-containing proteins and histone demethylation. Nat Rev Genet. 7:715–727.
Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. 2016. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv. doi: http://dx.doi.org/10.1101/071282.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.
Li H, *,1000 Genome Project Data Processing Subgroup, et al. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079.
Loose M. 2014. minoTour—a platform for real-time analysis and management of Oxford Nanopore MinION reads https://dx.doi.org/10.6084/m9.figshare. 1159099.v2.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.
Martinez J, et al. 2012. Influence of the virus LbFV and of *Wolbachia* in a host-parasitoid interaction. PLoS One 7:e35081.
Martinez J, Lepetit D, Ravallec M, Fleury F, Varaldi J. 2015. An additional heritable virus in the parasitic wasp *Leptopilina boulardi*: prevalence, transmission and phenotypic effects. J Gen Virol. 97:523–535.
McConkey GA, Martin HL, Bristow GC, Webster JP. 2013. Toxoplasma gondii infection and behaviour - location, location, location?. J Exp Biol. 216:113–119.
McVoy MA, Ramnarain D. 2000. Machinery to support genome segment inversion exists in a Herpesvirus which does not naturally contain invertible elements. J Virol. 74:4882–4887.
Moore J. 2013. An overview of parasite-induced behavioral alterations - and some lessons from bats. J Exp Biol. 216:11–17.
Mu J, et al. 2014. Live imaging of baculovirus infection of midgut epithelium cells: a functional assay of per os infectivity factors. J Gen Virol. 95:2531–2539.
Patot S, et al. 2010. Prevalence of a virus inducing behavioural manipulation near species range border. Mol Ecol. 19:2995–3007.
Patot S, Lepetit D, Charif D, Varaldi J, Fleury F. 2009. Molecular detection, penetrance, and transmission of an inherited virus responsible for behavioral manipulation of an insect parasitoid. Appl Environ Microbiol. 75:703–710.
Peng K, van Oers MM, Hu Z, van Lent JWM, Vlak JM. 2010. Baculovirus per os infectivity factors form a complex on the surface of occlusion-derived virus. J Virol. 84:9497–9504.
Piper DE, et al. 2015. The high resolution crystal structure of human LCAT. J Lipid Res. 56:1711–1719.
Renault S, Stasiak K, Federici B, Bigot Y. 2005. Commensal and mutualistic relationships of reoviruses with their parasitoid wasp hosts. J Insect Physiol. 51:137–148.
Rohrmann GF. 2014. Baculovirus molecular biology. 3rd ed. Bethesda (MD): National Center for Biotechnology Information (US).
Ruby JG, Bellare P, DeRisi JL. 2013. PRICE: software for the targeted assembly of components of (meta) genomic sequence data. G3&#58. Genes Genomes Genet. 3:865–880.
Shannon P, Markiel A, Ozier O, Baliga NS. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–2504.
Smith J. 2007. A gene's – eye view of symbiont transmission. Am Nat. 170:542–550.
Sovic I, et al. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 7:1–11.
Stamatakis A, et al. 2012. RAxML-Light: a tool for computing terabyte phylogenies. Bioinformatics 28:2064–2066.
Sugiura N, et al. 2013. Chondroitinase from baculovirus *Bombyx mori* nucleopolyhedrovirus and chondroitin sulfate from silkworm *Bombyx mori*. Glycobiology 23:1520–1530.
Sugiura N, Setoyama Y, Chiba M, Kimata K, Watanabe H. 2011. Baculovirus envelope protein ODV-E66 is a novel chondroitinase with distinct substrate specificity. J Biol Chem. 286:29026–29034.
Van Alphen JJ, Visser ME. 1990. Superparasitism as an adaptive strategy for insect parasitoids. Annu Rev Entomol. 35:59–79.
Van Houte S, Ros VID, van Oers MM. 2013. Walking with insects: molecular mechanisms behind parasitic manipulation of host behaviour. Mol Ecol. 22:3458–3475.
Van Lenteren JC, Bakker K. 1975. Discrimination between parasitised and unparasitised hosts in the parasitic wasp *Pseudeucoila bochei*: a matter of learning. Nature 254:417–419.
Varaldi J, et al. 2003. Infectious behavior in a parasitoid. Science 302:1930.
Varaldi J, Boulétreau M, Fleury F. 2005. Cost induced by viral particles manipulating superparasitism behaviour in the parasitoid *Leptopilina boulardi*. Parasitology 131:161–168.
Varaldi J, et al. 2006. Artifical transfer and morphological description of virus particles associated with superparasitism behaviour in a parasitoid wasp. J Insect Physiol. 52:1202–1212.
Zemskov EA, Kang W, Maeda S. 2000. Evidence for nucleic acid binding ability and nucleosome association of *Bombyx mori* nucleopolyhedrovirus BRO proteins. J Virol. 74:6784–6789.

Associate editor: Chantal Abergel