Article

Characterization of a Novel Mitovirus of the Sand Fly *Lutzomyia longipalpis* Using Genomic and Virus–Host Interaction Signatures

Paula Fonseca [1], **Flavia Ferreira** [2], **Felipe da Silva** [3], **Liliane Santana Oliveira** [4,5], **João Trindade Marques** [2,3,6], **Aristóteles Goes-Neto** [1,3], **Eric Aguiar** [3,7,*,†] and **Arthur Gruber** [4,5,8,*,†]

[1] Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 30270-901, Brazil; camargos.paulaluize@gmail.com (P.F.); arigoesneto@gmail.com (A.G-N.)

[2] Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 30270-901, Brazil; fvianaferreira@gmail.com (F.F.); jtm@ufmg.br (J.T.M.)

[3] Bioinformatics Postgraduate Program, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 30270-901, Brazil; felselva@gmail.com

[4] Bioinformatics Postgraduate Program, Universidade de São Paulo, São Paulo 05508-000, Brazil; liliane.sntn@gmail.com

[5] Department of Parasitology, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo 05508-000, Brazil

[6] CNRS UPR9022, Inserm U1257, Université de Strasbourg, 67084 Strasbourg, France

[7] Department of Biological Science (DCB), Center of Biotechnology and Genetics (CBG), State University of Santa Cruz (UESC), Rodovia Ilhéus-Itabuna km 16, Ilhéus 45652-900, Brazil

[8] European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany

* Correspondence: ericgdp@gmail.com (E.A.); argruber@usp.br (A.G.)

† Both corresponding authors contributed equally to this work.

Abstract: Hematophagous insects act as the major reservoirs of infectious agents due to their intimate contact with a large variety of vertebrate hosts. *Lutzomyia longipalpis* is the main vector of *Leishmania chagasi* in the New World, but its role as a host of viruses is poorly understood. In this work, *Lu. longipalpis* RNA libraries were subjected to progressive assembly using viral profile HMMs as seeds. A sequence phylogenetically related to fungal viruses of the genus *Mitovirus* was identified and this novel virus was named Lul-MV-1. The 2697-base genome presents a single gene coding for an RNA-directed RNA polymerase with an organellar genetic code. To determine the possible host of Lul-MV-1, we analyzed the molecular characteristics of the viral genome. Dinucleotide composition and codon usage showed profiles similar to mitochondrial DNA of invertebrate hosts. Also, the virus-derived small RNA profile was consistent with the activation of the siRNA pathway, with size distribution and 5′ base enrichment analogous to those observed in viruses of sand flies, reinforcing *Lu. longipalpis* as a putative host. Finally, RT-PCR of different insect pools and sequences of public *Lu. longipalpis* RNA libraries confirmed the high prevalence of Lul-MV-1. This is the first report of a mitovirus infecting an insect host.

Keywords: *Mitovirus*; positive-sense single-stranded RNA viruses; mitochondrial viruses; RNA interference; nucleotide frequency; codon usage; seed-driven progressive assembly; profile Hidden Markov Model

1. Introduction

Viruses are the most abundant biological entities in the biosphere, being found in every environment and infecting a wide range of organisms, such as plants, insects, mammals, and microorganisms [1–3]. Surveys to detect, identify, and characterize viral diversity are challenging due to the limited ability to isolate and grow viruses and their hosts in laboratory [4]. Furthermore, viruses do not have universally conserved sequences in their genomes that can be used as targets for PCR-based assays, such as the ribosomal genes

of prokaryotes and eukaryotes [5,6]. Finally, viruses present much higher evolutionary rates than prokaryotes and eukaryotes [7–10], which often implies that novel viruses are just too divergent to be detected by serological and molecular assays designed for specific known pathogens [6,11]. Metagenomics was classically defined as a sequence analysis method using samples containing multiple organisms [12]. With the advent of high-throughput sequencing platforms, metagenomics has greatly accelerated the pace of genome characterization and detection of viruses and hosts from environmental and clinical samples, without the need for isolation and prior cultivation [6,11,13,14]. Such an approach has allowed researchers to unveil viral diversity and virus–host interactions in many eukaryotic and prokaryotic organisms [15–17].

The relationship between viruses and their hosts is considered a coevolutionary process since viruses are obligate intracellular parasites and require the host's cellular machinery for protein synthesis. Also, viruses are subject to the same evolutionary pressures that shape the host genome composition and codon usage [18,19]. In fact, the coding regions of hosts and viruses tend to share common compositional features, such as dinucleotide composition and codon usage patterns [18–20]. Dinucleotide under- and over-representations are among the most studied and relevant of these patterns, and can be used to infer ecological functions and to classify viruses [21,22]. Nevertheless, dinucleotide composition can eventually be more related to the characteristics of the virus family than to the specific viral host species [23].

Hosts and their viruses are under constant adaptation and selective pressure, which is led by hosts developing new defense strategies and viruses developing new infection and counter-defense strategies [24]. One of the strategies developed by eukaryotic organisms against viral infections relies on the RNA interference (RNAi) pathways. RNAi pathways are mechanisms that induce silencing of self and non-self RNAs based on sequence-specific homology using small RNAs (sRNAs) [25]. In insects, it is well known the existence of three separate RNAi pathways- micro-RNA (miRNA), piwi-interacting RNA (piRNA), and small interference RNA (siRNA), with the latter being described as a hallmark of antiviral response in these organisms [26,27]. The siRNAs pathway is activated when double-stranded RNA is recognized and processed by the enzyme Dicer-2 into 19–23 nt long duplex of sRNAs. These sRNAs, are then loaded into Argonaute-2 to generate the small interferent RNA-induced silencing complex (siRISC) that produces virus-derived small RNAs, which in turn are used to find and cleave complementary RNAs. Interestingly, our group has shown that the size profiles of virus derived small RNAs are unique and distinctive, depending on the combination of host and virus species. Such feature can be used to determine the origin of viral sequences identified in metagenomic samples [25,27] and to differentiate between endogenous and exogenous viral sequences, which is a major problem in studies based solely on long RNA sequencing [25,26,28].

Recent studies have revealed that insects exhibit an extraordinary diversity and abundance of viruses [15]. Among invertebrates, insect vectors such as mosquitoes and phlebotomies have been extensively studied since they are associated with the transmission of several viral pathogens that threaten human health [29,30]. Sandflies are insects belonging to the order Diptera, subfamily Phlebotominae, which present hematophagous feeding habits. There are circa 900 species already described, and 70 of these species have been reported as potential vectors of *Leishmania* spp. with few others involved in the natural transmission of viruses, such as *Phlebovirus* (*Reoviridae* family) [31]. These insects are still capable of harboring other microorganisms, since they have contact with different environments and substrates [32–35]. This aspect can be especially relevant since many of these microorganisms that make up insect microbiota can also carry viruses.

Viruses of the genus *Mitovirus* were formerly classified [36] in the family *Narnaviridae*, together with the genus *Narnavirus* [37]. Both genera show distinct subcellular localizations, comprise capsidless viruses with a monopartite positive sense single-stranded RNA genome of 2.3–2.9 kb, and present a single gene encoding an RNA-dependent RNA polymerase (RdRp). In the case of *Mitovirus*, the RdRp gene presents a mitochondrial-type

codon usage, with UGA coding for tryptophan. In the current International Committee on Taxonomy of Viruses (ICTV) report, as of March of 2020—Master Species List #35 [38], *Narnaviridae* family, containing the genus *Narnavirus*, was included in the *Wolframvirales* order, whereas genus *Mitovirus* now belongs to a newly created *Mitoviridae* family, order *Crippavirales*. Both orders are currently members of the *Lenarviricota* phylum, which also comprises *Leviviridae* (order *Levivirales*), a family of positive-sense single-stranded RNA bacteriophages that may have originated narnaviruses, mitoviruses, and ourmiaviruses [39]. Mitoviruses have been identified in many fungal hosts, such as *Entomophthora muscae* and *Fusarium boothii* [40,41] and narnaviruses in invertebrates such as insects and other arthropods [15], but, differently from mitoviruses, their replication occurs in the cytoplasm of the fungal hosts [37,42,43]. In addition to *Narnaviridae* family, fungi can also be infected with other viruses (mycoviruses), which replicate in the cytoplasm of the host cells [44]. Similar to *Narnaviriridae* and *Mitoviridae*, the *Botourmiaviridae* family is composed of viruses infecting the cytoplasm of plant cells and, unlike these two families, their genome is composed of three monocistronic segments [45]. Nevertheless, some studies have identified members of the *Ourmiaviridae* family infecting the cytoplasm of filamentous fungi [46].

Although there is an increasing number of reports uncovering the diversity of viruses circulating in insects, they are mainly restricted to nucleic acid sequencing, with no additional biological characterization, thus restricting the ability to determine the origin of the viral sequences [2,15]. Also, most of the viral surveys reported in the literature rely on the use of conventional pairwise similarity searches, which often yield no identification of the sequences found [47]. It has been demonstrated that pairwise similarity searches are effective in detecting relatively close homologs, but fail to identify distantly related sequences [48]. Conversely, similarity methods using sequence profiles are able to detect remote homologs with much higher sensitivity [49]. Profile Hidden Markov Models (profile HMMs) are probabilistic models built from multiple sequence alignments that cover the variability of residues in all positions, including indels and inserts [6]. Such models have been increasingly used in viral classification and discovery [50–53].

An additional challenge to detect novel viruses from metagenomics samples is to assemble large metagenomic datasets composed of an unknown number of different organisms. An alternative method for DNA assembly was described by our group and implemented on GenSeed-HMM [54], a program that uses profile HMMs as seeds for targeted progressive assembly. Such approach can be used in many applications, including viral discovery [6,54].

In this work, we investigate the diversity of viruses circulating in the sandfly *Lutzomyia longipalpis*, the most important vector of *Leishmania chagasi* in the New World. For this goal, we use three innovative methods: (1) profile HMMs to interrogate public long RNA sequencing data, (2) progressive assembly using profile HMMs as seeds, and (3) small RNA profiles to differentiate exogenous from endogenous viral sequences. Using this integrated approach, we identify and describe Lul-MV-1 (*Lutzomyia longipalpis mitovirus 1*), the first mitovirus found to infect the mitochondria of an insect host.

2. Materials and Methods

2.1. Acquisition and Processing of RNA Libraries

Lu. longipalpis public libraries of long and small RNAs were downloaded from the NCBI Sequence Read Archive (SRA) repository (https://www.ncbi.nlm.nih.gov/sra). Accession numbers of the analyzed libraries are listed in Table S1. In total, six small RNA (sRNA) libraries and two long RNA (lRNA) libraries were used in this study. Libraries were submitted to quality end-trimming and adapter removal. Sequences presenting < 80% of bases with Phred quality below 20 or a length shorter than 20 bases were discarded. The remaining reads were used in further analyses.

2.2. Profile HMM Screening and Progressive Assembly

Long RNA reads of *Lu. longipalpis* public libraries (Table S1) were used for viral sequence detection and reconstruction. A subset of 506 profile HMMs was manually built from the vFAM database [50]. These models were chosen based on their unequivocal functional annotation and for representing virus-specific proteins or sequences distantly related to prokaryotic or eukaryotic orthologs. We used HMM-Prospector (https://github.com/gruberlab/hmmprospector [accessed on 20 December 2020]), a Perl program to screen the profile HMMs against 6-frame translated versions of the long RNA datasets. HMM-Prospector uses hmmsearch program from HMMER package v. 3.1 [55] to run similarity searches and then processes the results, generating tabular files with qualitative and quantitative results. Profile HMMs detecting the highest numbers of significant hits (score > 30 and/or e-value < 1×10^{-5}) were used as seeds for GenSeed-HMM [54], a tool for seed-driven progressive assembly. The reconstructed sequences were submitted to sequence similarity searches using BLASTX [56] against the non-redundant (nr) NCBI database. The programs Artemis (v. 16.0) [57] and InterProScan (version 5.36–75.0) [58] were used to detect open reading frames (ORFs) and conserved domains, respectively. Hits with e-values smaller than 1×10^{-5} for nucleotide comparison or 1×10^{-3} for protein comparison were considered significant. Viral genomic segments were classified as described [59].

2.3. Phylogenetic Analysis

A dataset composed of public protein sequences (Table S2) related to mitoviruses, narnaviruses, and ourmiaviruses/ourmia-like viruses was constructed and submitted to a multiple sequence alignment with MUSCLE [60]. Phylogenetic reconstruction was performed by using IQ-TREE version 1.6.11 [61] with ModelFinder [62] to determine the model that minimizes the BIC (Bayesian Information Criterion) score. Node support values were determined using 1000 pseudoreplicates with the ultrafast bootstrap approximation (UF-Boot) method [63]. The obtained trees were visualized and edited with Dendroscope [64].

2.4. Analysis of Small RNA Libraries

The pre-processing of sRNA libraries was performed as described [25]. Briefly, small RNAs reads were mapped against assembled contigs or viral genomes using Bowtie [65] allowing one mismatch. Small RNA size profile was calculated as the frequency of each small RNA read size mapped on the reference genome or contig sequence considering each polarity separately. We used a Z-score to normalize the small RNA size profile and to plot heatmaps for each sequence using R language (version 3.0.3) with gplots package (version 2.16.0). Pearson correlation with a confidence interval >95% of the Z-score values were computed to evaluate the relationships between the small RNA profiles from different contigs or reference genomes. The profile similarity was assessed using hierarchical clustering with UPGMA as the linkage criterion. Groups of sequences with more than one element with at least 0.8 of Pearson correlation between each other were assigned to clusters. Small RNA size profile, 5′ base enrichment, density of coverage and additional data analysis were evaluated using in-house Python, Perl, and R scripts. Statistics of 5′ base enrichment was calculated as described [66]. Similarities between small RNA size distributions were defined using hierarchical clustering with *K-means* as the linkage criterion in R using corrplot package [67]. Empirical cumulative frequency of small RNA size distribution was computed and compared using *ecdf* function built-in R software where the Kolmogorov–Smirnov test was used to determine statistical significance.

2.5. Dinucleotide and Codon Usage Analyses

Dinucleotide frequencies and codon usage were calculated using programs from the EMBOSS package (version 6.6.0) [68]. First, we used the program extractfeat to extract the coding sequences (CDS) from GenBank files and to store the data in FASTA format. Next, we used compseq to calculate the composition of unique 2-mer words in all frames to

determine the dinucleotide expected/observed frequencies. Finally, the cusp program was used to generate codon usage tables containing the number of codons per 1000 bases, given the input sequence and the proportion of usage of each codon among its redundant set. The correlation between virus and host frequencies was calculated using the Pearson correlation test. Dinucleotide frequency was plotted using the R package corrplot, which grouped elements into clusters based on the results of the Pearson correlation test with a threshold above 0.8. Codon usage values were plotted as a heatmap with groups containing elements with mutual Pearson's correlation coefficients of at least 0.8. The viral and mitochondrial genomes of fungal and insect hosts analyzed in this study are listed in Table S3.

2.6. Amplification and Sanger Sequencing

To confirm the presence of the virus found in *Lu. longipalpis*, we used eight pools containing 10 sandflies per pool, collected from a colony originally started from individuals collected in Teresina, Brazil, and maintained at the Laboratory of Physiology of Hematophagous Insects (Department of Parasitology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais). Total RNA was extracted from pools of 5 insects each using Trizol reagent according to the manufacturer's protocol (Invitrogen—Thermo Fisher Scientific Inc., Waltham, MA, USA). Total RNA (1 μg) was reverse transcribed using 250 ng of random primers specific primers per reaction. The resulting cDNA was used as template for PCR reaction containing primers designed to amplify a product of 509 bp. Primer sequences are listed in Table S4. Conventional PCR was performed using 1.5 μL of each designed primer (10 pmol/μL), 200 ng of cDNA or DNA and *Taq* DNA polymerase (Invitrogen, Thermo Fisher Scientific Inc.). PCR products were cleaned up with EDTA 125 mM precipitation protocol and sequenced using Sanger technology.

2.7. Analyses of Public Libraries

DNA and long and small RNA libraries were obtained from the NCBI SRA repository, listed with accession numbers in Table S1. To estimate and analyze the abundance of the putative viruses, each library was compared to the viral and mitochondrial genomes using Bowtie2 [65]. The result was normalized by Reads Per Million (RPM) and plotted on a bar graph using the R program with the ggplot2 package [69].

3. Results

3.1. Identification of Viral Sequences in Lu. longipalpis Datasets

In a first attempt to detect possible viral sequences in *Lu. longipalpis*, we tested a set of 506 profile HMMs selected from the vFam database against two sequencing datasets of lRNA data, totaling 48 million reads. This strategy allowed us to select 18 different profile HMMs (Table S5), which were used for progressive assembly, leading to the identification of 288 putative viral sequences. We observed a high abundance of contigs related to reverse transcriptase and integrase, in agreement with the high number of transposable and retroviral elements found in insect genomes [70] (Figure S1; Table S6). In addition, we identified 18 contigs (vFam 561 and vFam 1529) derived from nucleoprotein N gene from *Rhabdovirus* (Table S6), commonly found integrated in the genomes of many eukaryotes [71,72]. The other vFam models represented less than 10.1% of the identified contigs. Since the presence of symmetrical small RNAs of 20–23 nt, derived from viral sequences, is an indicative of activation of the siRNA pathway during viral replication [25], we decided to investigate the small RNA size distribution of all contigs reconstructed by profile HMM-seeded progressive assembly. From 288 assembled contigs representing putative viral sequences, only one sequence, reconstructed from vFam_571 model (Table S6), presented a size distribution of small RNAs consistent with the activation of the siRNA pathway— symmetrical peak at position 21, derived from both strands without 5′ base preference (Figure 1).
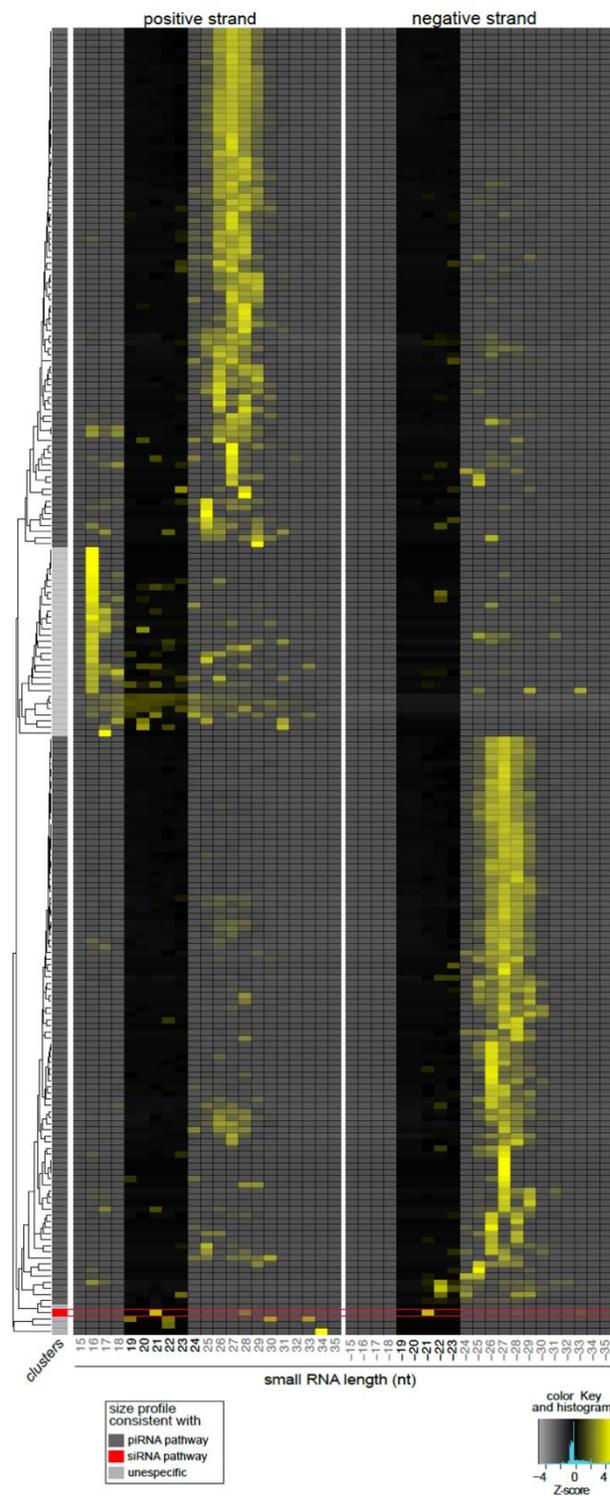
Figure 1. Small RNA size distribution of putative viral sequences assembled from sandfly RNA libraries. Clustering is shown as a heatmap based on Pearson correlation of the sRNA profile. Clusters are defined by a Pearson correlation above 0.8. Viral contigs were classified according to the small RNA size distribution expected for siRNA and piRNA populations in insects. Shaded columns represent small RNAs with length between 19 and 23 nt.

The viral sequence assembled from vFam_571 presented a total length of 2697 nt with a GC content of 30.26%, corresponding to a monopartite ssRNA(+) genome. BLASTN searches against public databases showed no significant similarity. Using the universal

genetic code, only short open reading frames (ORFs) were observed. BLASTX searches against the nr database revealed some similarity to RdRp of narna-like viruses such as the Wenling narna-like virus 9 (accession code YP_009337200) [15], which use UGA to encode tryptophan instead of signaling translation termination. In fact, when we switched to a mitochondrial genetic code, we found a long CDS coding for an 804-aa protein. InterProScan search of the protein sequence confirmed a positive identification for a mitoviral RdRp (InterPro entry IPR008686) and the presence of the Pfam domain PF05919. This nucleotide sequence corresponds to an almost complete genome sequence of a novel virus, hereinafter referred to as Lul-MV-1, the *Lutzomyia longipalpis* mitovirus 1.

3.2. Phylogenetic and Genome-Based Characterization of Lu. longipalpis Mitovirus 1

We performed a phylogenetic analysis with representatives of the current families *Mitoviridae* and *Narnaviridae* [38], who were formerly part of a common family [73], including some viral prototypic species ratified by the ICTV. We also included members of the genus *Ourmiavirus* (*Botourmiaviridae* family) and some unclassified viruses. Finally, two enterobacteria phages of the *Leviviridae* family were used as an outgroup, since these viruses are close relatives to *Mitovirus* and *Narnavirus* [74]. In addition, it has been proposed that these genera might have been evolved from *Leviviridae*-ancestors that infected bacterial endosymbionts, some of which may have generated mitochondria [39,45,75]. A phylogenetic reconstruction using RdRp sequences (Figure 2) revealed that Lul-MV-1 is closely related to the *Plasmopara viticola lesion-associated mitovirus 56* (QIR30279), a fungal mitovirus found in grapevine [76], and to a slightly lesser extent to the *Wenling narna-like virus 9* (YP_009337200), identified in crustaceans [15]. The narnaviruses constituted a sister clade composed of viruses infecting either fungal or insect hosts. The monophyly of the genera *Mitovirus* and *Narnavirus* was clearly supported by our analysis. Some viruses such as the *Grapevine associated narnavirus 1* (accession code CEZ26304), *Wenling narna-like virus 9* (YP_009337200), and the *Shahe narna-like virus 6* (APG77166), originally classified as narnaviruses, are members of the genus *Mitovirus* according to our analysis (Figure 2). Finally, members of the genus *Ourmiavirus* and some unclassified viruses constituted a sister clade to narnaviruses. In addition to the prototypical plant viruses such as the Cassava virus C (YP_003104770), *Ourmia melon virus* (YP_002019757) and *Epirus cherry virus* (YP_002019754), some recently described narna-like viruses are in fact ourmia-like viruses. Similarly, the *Aspergillus fumigatus mitovirus 1* (AXE72932), originally classified as a mitovirus [77], is clearly misclassified, and also belongs to the *Ourmiavirus*/ourmia-like clade. This group presents a high divergence among their members and it is possible that larger taxa samplings may indicate in the future that it is polyphyletic indeed.

3.3. Comparative Analysis of Structural and Compositional Features

In addition to the phylogenetic reconstruction, we performed a comparative analysis of the *Mitovirus* and *Narnavirus* genera according to the size of the viral genome and protein sequences (Figure 3). The coding sequence of the genus *Mitovirus* presents an average size of 2000–2500 nt and an ORF (Open Reading Frame) coding for an RdRp with a maximum size of 900 aa residues. Conversely, the *Narnavirus* genus shows an average genome size of 2500–3000 nt and an RdRp ORF with a maximum length of up to 1200 aa residues. Some narnaviruses also present ambigrammatic sequences, characterized by an additional large ORF coded in the reverse strand of the genome [78,79].
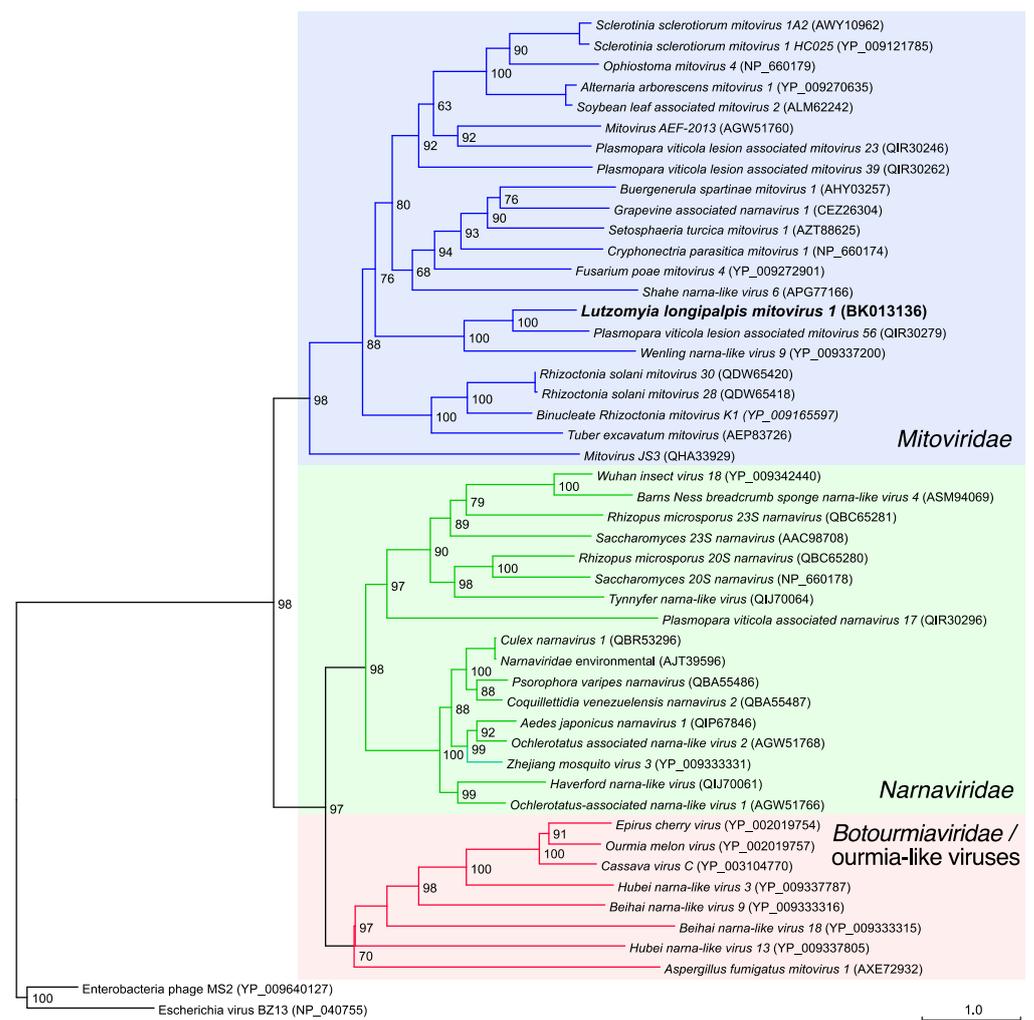
Figure 2. Phylogenetic analysis. Phylogenetic tree inferred by maximum likelihood using full-length amino acid sequences of the RNA-dependent RNA polymerase (RdRp). Phylogenetic reconstruction was performed with IQ-TREE using the evolutionary model Blosum62 + F + R6 and 1000 bootstrap pseudoreplicates. The tree was rooted using two sequences of ssRNA bacteriophages belonging to *Leviviridae* family. Colored clades correspond to the familis *Mitoviridae* (blue) and *Narnaviridae* (green), and to *Botourmiaviridae*/ourmia-like viruses (red).

To characterize and propose a probable host of Lul-MV-1, we analyzed some intrinsic features of the viral genome. First, we compared the dinucleotide frequencies of Lul-MV-1 with those of mitochondrial genomes of fungi, mosquitoes (Culicidae) and phlebotomines (Phlebotominae). Lul-MV-1 presents a dinucleotide usage profile similar to the profiles observed in insects, but not to those of fungi (Figure 4). In fact, Lul-MV-1 presents a highly biased composition of the dinucleotides CC and GG and a mid-low GC bias, resembling the frequencies of insect mitochondria. A comparative clustering analysis of dinucleotide frequencies grouped Lul-MV-1 together with mitoviruses that infect fungi, whereas narnaviruses of fungi and insects were clearly grouped in another cluster (Figure S2). This result corroborates the genus assignment observed for Lul-MV-1 in the phylogenetic analysis (Figure 2). Finally, mitochondrial genomes of eukaryotic hosts were clustered into two closely related groups comprising fungi and insects, respectively (Figure S2).
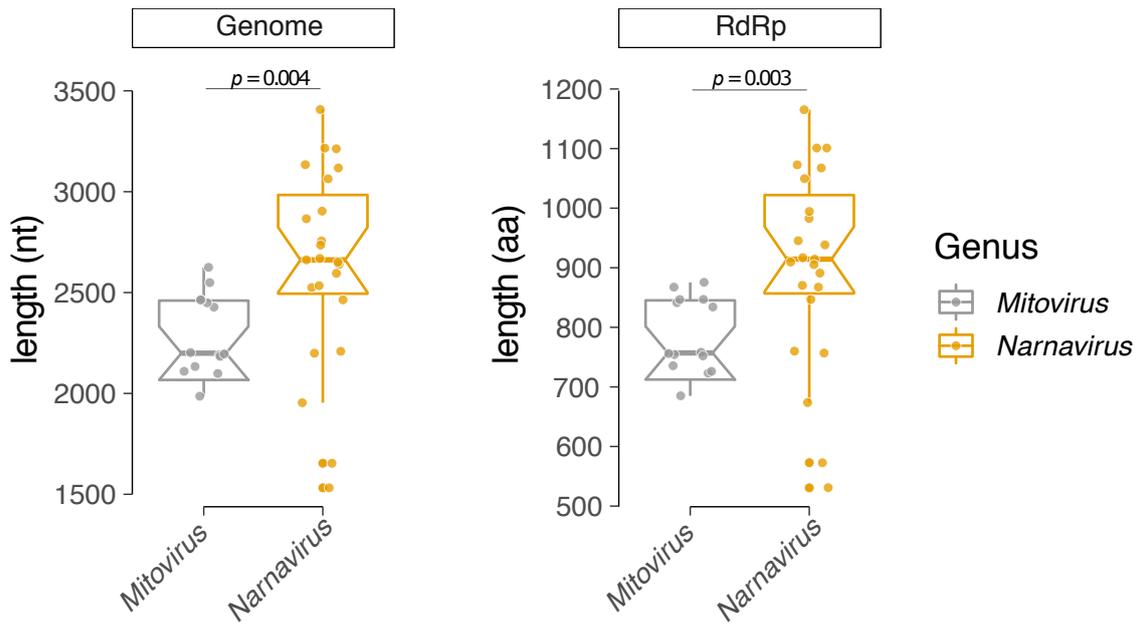
Figure 3. Comparative genome length. Comparison of genome and RNA-dependent RNA polymerase (RdRp) lengths of viruses of the genera *Narnavirus* and *Mitovirus*.
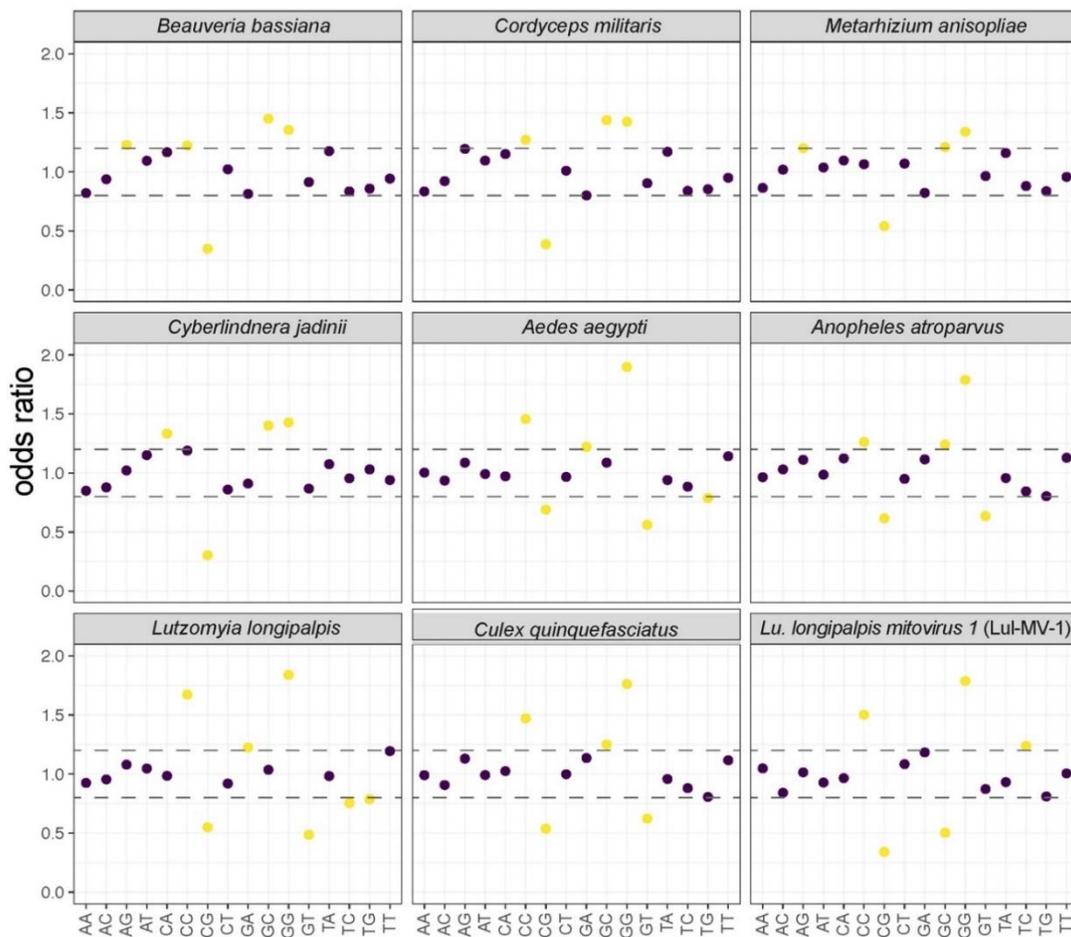

Figure 4. Dinucleotide frequency of Lul-MV-1 and nuclear genomes of fungi that commonly infect insects. Yellow circles represent dinucleotide odds ratios which are highly biased regarding the expected frequencies. Purple circles refer to dinucleotide odds ratios within the unbiased region (delimited by dashed lines). Dashed lines indicate cutoffs values of 0.78–1.25.

Since di- and trinucleotides patterns can shape codon usage frequency in viruses [80], we also compared the codon usage profile of Lul-MV-1 with the profiles of some mitoviruses and mitochondrial genomes of fungi and insects, which in turn are the known or putative hosts of some viruses. Using a hierarchical clustering of these codon usage profiles, organisms presenting Pearson correlation values of at least 0.8 were grouped into the same clusters. From the six observed groups, clusters 1–3 are exclusively composed of mitoviruses infecting fungal hosts (Figure 5). Clusters 4 and 5 are composed of mitochondrial genomes of insects and fungi, respectively, with exception of Lul-MV-1, which is present in cluster 4, closely related to the mitochondrial genome of *Lu. longipalpis*. This result suggests that Lul-MV-1 is highly adapted to the mitochondrial codon usage of its putative host. The *Shahe narna-like virus 6* [15], characterized from a metagenomic dataset of crustaceans, showed a codon usage frequency closely related to the mitochondrial genome of the fungus *Plasmopara viticola*. Nevertheless, the *Plasmopara viticola associated mitovirus 56* presented a profile distantly related to its host, but closely related to other fungal mitoviruses. Finally, the *Wenling narna-like virus 7*, another virus detected in a metagenomic sample derived from crustaceans [15], showed a codon usage frequency unrelated to any of the tested organisms.
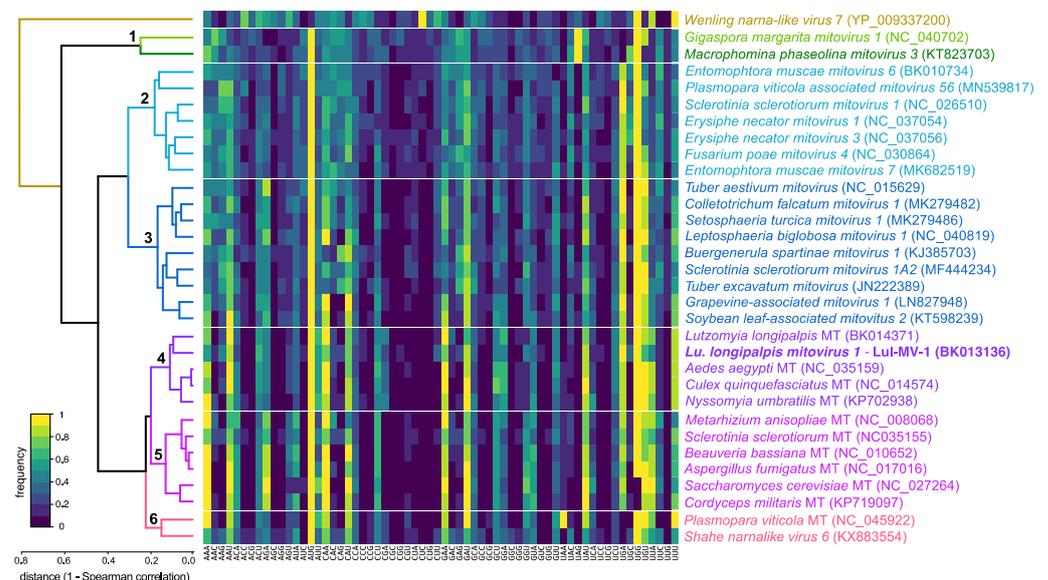

Figure 5. Relationship between codon-usage of mitochondrial viruses and mitochondrial genomes (MT) of putative hosts. Hierarchical clustering of codon usage profiles of viruses and mitochondria of their respective hosts. Clusters were defined by Pearson correlation above 0.8. The numbers indicate the main clades referred in the text.

To better understand these results, we decided to restrict and deepen the analysis to UGA/UGG codons. Mitochondrial genomes often use a genetic code with UGA coding for tryptophan (Trp) rather than acting as stop codon. Thus, we chose a set of viral and mitochondrial genomes and determined the absolute counts and relative usage frequencies of these codons (Table S7). In all cases where only one single UAG codon was counted, manual curation revealed that it was in fact acting as a stop codon at the end of coding sequence, rather than coding for tryptophan. In the case of the sampled narnaviruses, we did not observe the use of UGA(Trp) codons, either in fungal or insect viruses. More interestingly, this pattern was also seen for the hypothetical genes present in the reverse frame of some of these viral genomes. Conversely, the mitochondrial genomes of the corresponding fungal and insect hosts used almost exclusively UGA(Trp). This discrepant codon usage was highly correlated with the AT (adenine-thymine) content, with hosts' mitochondria showing very high AT content, in the range of 70–80%, whereas the viruses presented much lower values, around 40%.

Mitoviruses, which are known to locate exclusively in the mitochondria of their hosts, showed a variable ratio of UGA/UGG codons (Table S7), with a relatively close correlation with the ratios observed in the respective host's mitochondrion. For instance, *Cryphonectria mitovirus 1* and *Cryphonectria parasitica* mitochondrion used UGA in 52.9 and 82.1% of the tryptophan codons. In another case, *Sclerotinia sclerotiorum mitovirus 1* and *Sclerotinia sclerotiorum mitovirus 1-A2* used UGA(Trp) in 76.9 and 66.7%, respectively, with the host's mitochondrion showing a usage of 81.2%. In *Ophiostoma novo-ulmi*, another ascomycete fungus, the mitochondrion and *Ophiostoma mitovirus 4* used UGA(Trp) in 94.1 and 84.6%, respectively. A remarkable distinct result was obtained for *Plasmopara viticola*, an oomycete whose mitochondrion exclusively used UGG to encode tryptophan, while the *Plasmopara viticola associated mitovirus 56* showed a 50% use of UGA and UGG codons. In the case of Lul-MV-1, the mitovirus presented 73.3% of UGA(Trp), with the *Lutzomiya longipalpis* mitochondrion using 99%. Two other mitoviruses, found in crustacean metagenomic samples, the UGA(Trp) utilization showed an extreme variation, with the *Wenling narna-like virus 9* using 99% of UGA(Trp) and the *Shahe narna-like virus* exclusively using UGG. When analyzing the AT content, unlike narnaviruses, mitoviruses showed a relatively high correlation with their hosts' mitochondria.

In the case of the ourmiaviruses and ourmia-like viruses, we observed no use of UGA(Trp) at all. While *Aspergillus fumigatus'* mitochondrion used UGA(Trp) in 93.2% of the tryptophan codons, the *Aspergillus fumigatus mitovirus 1* (which is not a mitovirus indeed—see Figure 2) used UGA only in one single occurrence, as a stop codon. Also, the virus showed a much lower AT content than the host's mitochondria. Finally, both *Escherichia* phages belonging to the *Leviviridae* family shared with their host the use of a standard genetic code with UGA codon signaling translation end and showed very similar AT content values.

3.4. Lul-MV-1 Is Targeted by the Lu. longipalpis siRNA Pathway

Previous works from our group have shown that the virus-derived small RNAs (vsRNAs) present some features that are host-virus specific and can be used to classify viral sequences [27]. Therefore, we assessed the characteristics of the vsRNAs and compared them to the profiles of fungi-infecting viruses and to the *Vesicular stomatitis virus* (VSV), another virus that naturally infects *Lu. longipalpis* [34]. We observed that the small RNAs derived from Lul-MV-1 showed size distribution and 5′ base enrichment (21 nt symmetrical peak and absence of 5′ base enrichment) that are distinguishable from those observed for fungal viruses (20–22nt symmetrical peak with 5′ base preference for Uracil) (Figure 6A,B). In agreement with this result, an analysis of cumulative frequency, based on small RNA size distribution, showed that Lul-MV-1 presents a profile noticeably different from those observed in fungal viruses. Furthermore, there was no significant difference between the profiles of Lul-MV-1 and VSV (Figure 6C). This body of evidence strongly suggests that Lul-MV-1 is likely to infect the phlebotomine mitochondria, rather than the mitochondria of a putative fungal host.

The sequencing depth and coverage of sRNAs derived from the putative virus were also analyzed (Figure S3). We observed small RNAs mapping across the entire genome of Lul-MV-1 on both positive and negative strands with a similar pattern. A homogenous and symmetrical coverage of the viral genome is a typical signature of the siRNA pathway that is triggered by a dsRNA precursor, resulting in virus-derived small RNAs [27,81]. In fact, the Lul-MV-1-derived small RNA profile is very similar to the antiviral siRNA response observed in *Lutzomyia* and *Drosophila* [25,34,66].
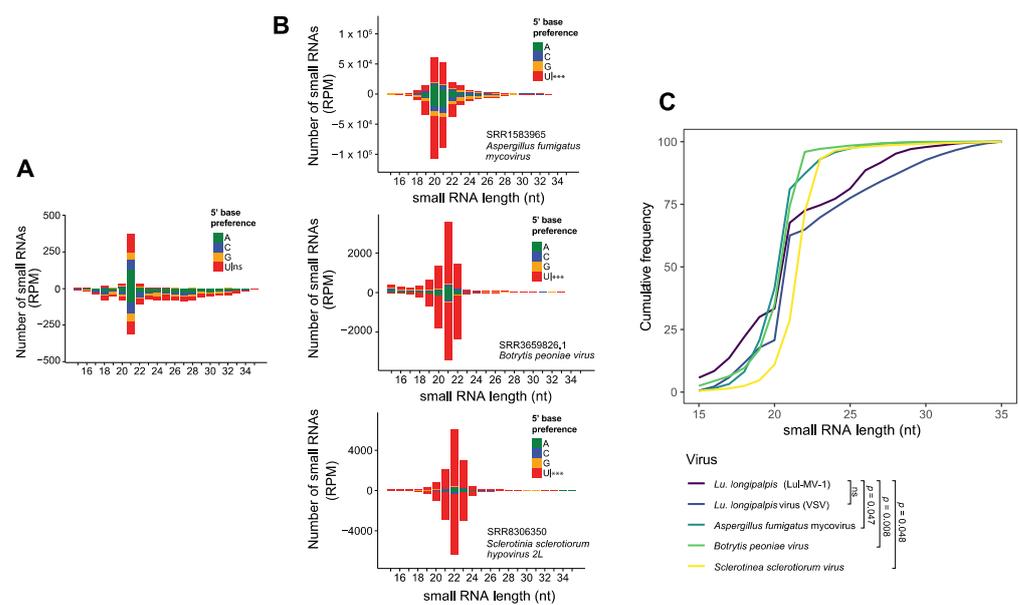
Figure 6. Characterization of small RNAs derived from *Lu. longipalpis* and fungi-related viruses. (**A**) Small RNA size profiles and 5′ base preference of small RNAs derived from Lul-MV-1 and (**B**) fungi-related viruses (from top to bottom: *Aspergillus fumigatus* mycovirus, *Botrytis paeoniae* virus and *Sclerotinia sclerotiorum* hypovirus). 5′ base preferences of small RNAs are indicated by color. Significant differences are also indicated. (**C**) Cumulative frequency according to the size distribution of small RNAs varying from 15 and 35 nt, derived from viruses infecting *Lutzomyia* (Lul-MV-1 and VSV) and fungi. Statistical significance among cumulative frequencies was determined using the Kolmogorov-Smirnov test.

3.5. Prevalence of Lul-MV-1 in Lu. longipalpis Colony and Public Datasets

To confirm the sequence and presence of Lul-MV-1 in *Lu. longipalpis*, we performed amplification by RT-PCR using pool samples derived from the same laboratory colonies used to obtain the small RNA libraries. We observed amplification of the Lul-MV-1 in seven out of the eight tested pools of *Lu. longipalpis*, indicating that this virus is present in high prevalence in *Lu. longipalpis* individuals (Figure 7A). Sanger sequencing of the PCR product confirmed that the Lul-MV-1 genome assembled in this work is 95% similar to the virus found in laboratory colonies in Brazil (Figure S4). This small divergence likely reflects the natural variation in viruses infecting different sandfly populations.

Since we detected the Lul-MV-1 in public libraries derived from sandfly colonies maintained in Cambridge, UK, and Brazil, we decided to investigate other sequencing datasets available in public databases. In total, we assessed 15 other RNA libraries and detected virus-derived reads in the majority of the samples, with exception of a small RNA library derived from *Lu. longipalpis* LULO cells (Figure 7B). Interestingly, embryo libraries showed large abundance of viral sequences, in some cases showing higher counts than the mitochondrial reads (Figure 7B). Concluding, our results indicate that Lul-MV-1 is highly prevalent in *Lu. longipalpis* populations.

To verify whether Lul-MV-1 represents a viral element integrated into the host genome, besides evaluation of *Lu. longipalpis* genome, we also interrogated public sequencing datasets derived from the same NCBI project from which the RNA libraries were extracted. As a positive control, we analyzed both the sequence of Lul-MV-1 and the mitochondrial genome of the sand fly. We observed a considerable number of reads derived from the mitochondrial genome, but a complete absence of reads derived from the viral genome, suggesting that the virus does not have an DNA intermediate form and neither represents an endogenous viral element integrated on the host genome (Figure 7B).
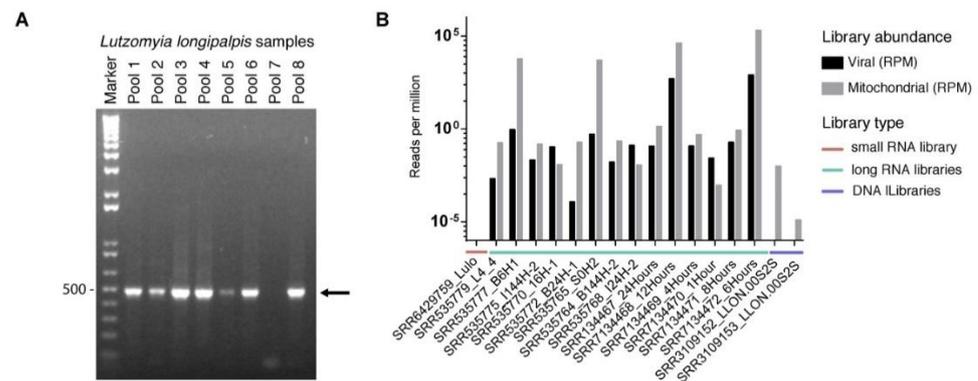
Figure 7. Detection of Lul-MV-1. (**A**) RT-PCR amplification of a Lul-MV-1 target using *Lu. longipalpis* RNA from different pools of individuals derived from the same colony used to prepare the small RNA deep sequencing libraries. The arrow indicates the amplification product. (**B**) Reads from different RNA and DNA public libraries of *Lu. longipalpis* mapped onto the genome of Lul-MV-1 with Bowtie2, quantified and resulting numbers normalized by Reads Per Million (RPM).

4. Discussion

Sand flies (subfamily Phlebotominae) are ubiquitous crepuscular-nocturn insects, found in all continents in both rural and urban areas. In the New World, *Lu. longipalpis* is the most important phlebotomine vector of *Leishmania chagasi*, the causing agent of human visceral leishmaniasis. Both males and females feed on sugar sources, but females are anautogenous and must ingest blood to provide protein substrates for egg-maturation and oviposition. This blood meal is obtained from a variety of mammals and birds, contributing for these insects to become a major primary reservoir in which viruses belonging to several families [31,82] can replicate and be transmitted across different host species [31,83]. Some of these viruses are restricted to insects and their role in the biology of these hosts is often poorly understood.

In this work, we described the genome of Lul-MV-1, a novel virus found in *Lu. longipalpis* RNA samples. We used profile HMMs together with GenSeed-HMM [54] to select virus-specific reads and perform a target-specific progressive assembly. The reconstructed genome revealed a single ORF coding for an RdRp, where the UGA codon is used for tryptophan instead of acting as a stop codon, a characteristic often seen in organelles such as mitochondria. A phylogenetic reconstruction, using maximum likelihood as the optimality criterion, positioned LuL-MV-1 in a monophyletic clade of viruses of the genus *Mitovirus*. Interestingly, our phylogenetic analysis (Figure 2) also showed a relatively close relationship with two viruses infecting invertebrate hosts, the *Wenling narna-like virus 9* (YP_009337200) and the *Shahe narna-like virus 6* (APG77166), both found in crustacean samples [15]. Unfortunately, the original samples were composed of a pool of sources and no specific hosts were assigned in the report.

In recent years, the increase of environmental metagenomic studies has provided the description and identification of virus sequences of the former *Narnaviridae* family (comprising both narnaviruses and mitoviruses—actually classified into distinct families) many organism hosts, such as invertebrates, fungi, plants and mammals [15,35,37,41,84–86]. However, it is still uncertain whether these viruses infect fungal and protist symbionts or organisms of the regular microbiota [78,79,84,87].

Narnaviruses have been reported to infect fungi, plants and dipteran insects [88,89]. The prototypic species of the genus *Narnavirus* were originally described infecting the cytoplasm of the yeast *Saccharomyces cerevisiae* [37]. In addition to the RdRp gene, some narnaviruses also contain a reverse-frame ORF. For example, the *Aedes japonicus narnavirus 1* presents a genome of 2069 bases containing two ORFs, one in the positive-sense strand coding for RdRp, and an additional negative-frame ORF. Ambisense coding strategy has been studied in narnaviruses of insects, suggesting that both ORFs could enable replica-

tion in the hosts. Reverse-frame ORFs are characterized by the avoidance of CUA, UUA, and UCA codons, which are the reverse complements of stop codons [79], a finding that suggests that these putative genes are active, but their biological function still remains unknown [78]. Nevertheless, is still unclear how this reverse-frame ORF would be translated, since eukaryotic translation depends on an initiation site close to 5′ ends of transcripts in the positive sense [79]. In the case of Lul-MV-1, no ambisense ORF was found, a feature that corroborates its classification within the *Mitovirus* genus. Sequences related to narnaviruses were also found in two samples of the mosquito *Culex pipiens* and phylogenetic analyses revealed that these sequences are closer to other narnaviruses associated to mosquitoes than to fungal narnaviruses [88].

Mitoviruses have been identified only in fungal and plant mitochondria. Mitovirus-like sequences closely related to fungal viruses, derived from a specific branch, were detected as endogenous elements integrated in plant mitochondrial genomes, and pathogenic fungi were raised as potential source of horizontal transfer [75]. A survey of transcriptomes of ten distinct plant revealed 20 complete sequences of mitoviruses and some results suggested that genuine plant mitoviruses may have originated endogenized mitovirus found in plants [84]. Since most of the mitoviruses have been typically found in fungal hosts, we initially assumed that Lul-MV-1 was infecting a fungus of the regular microbiota of *Lu. longipalpis* or, alternatively, that the source was a fungal contaminant. We used different analyses to confirm that Lul-MV-1 infects the insect's mitochondria, rather than the mitochondria of a fungal host. Dinucleotide composition represents one of the host adaptation mechanisms that influence virus codon usage [80]. As part of the virus–host adaptation process, dinucleotide composition and codon usage tend to have similar frequencies between viruses and their hosts [19]. In fact, if these compositional features are not optimized for the host, mRNA stability and protein synthesis are negatively impacted, reducing viral fitness and multiplication [80,90]. According to our results (Figure 4), Lul-MV-1 shows a dinucleotide frequency that is dissimilar to the composition of fungi but resembles that of insect hosts. Our codon usage analysis (Figure 5) showed that viruses infecting fungi showed closely related codon usage profiles, but they were more distantly related to their hosts' mitochondria. Conversely, Lul-MV-1 showed a codon usage profile closely related to insects and especially to *Lu. longipalpis*. This result shows a remarkable virus–host adaptation and points to this dipteran as the putative host of Lul-MV-1, especially considering the high prevalence and abundance of its viral sequence in public datasets derived from different sources of *Lu. longipalpis*.

A biological correlation of codon usage fitting by mycoviruses and fungal host virulence was reported in *Aspergillus* spp. [91]. Mycoviruses causing hypervirulence in fungi have an increased content of C or G at the third position, whereas viruses that do not alter the fungal host virulence do not share similar codon usage patterns, suggesting that mycovirus-mediated modulation of the host is dependent on the similar codon usages, specifically in the third position of the codons. Similar results were also observed in other systems [92], where viruses presenting codon usage biases similar to their hosts can impair translational efficiency and therefore reduce host fitness. In contrast, natural hosts, infected with viruses with dissimilar codon usage, present no changes in protein translation or fitness.

Many fungal mitoviruses show a large utilization of UGA(Trp) codons, resembling the codon usage of the respective mitochondria [41,90]. A comprehensive survey of codon usage profiles in fungal mitochondria revealed that UGA(Trp) are rarely used in many organisms [90]. Another comparative study of fungal mitogenomes also showed variability in terms of genetic code, comprising the use of genetic codes 1, 4 and 16 [93]. These results suggest that viruses mimicking the mitochondrial genetic code could have a more efficient use of the translational machinery of the organelle. According to Nibert [90], the exclusion of UGA(Trp) codons in some viruses would just reflect the scarcity of these codons in the mitochondria of their specific hosts. In agreement with this hypothesis, our results show that mitoviruses present an overall AT content and UGA/UGG usage ratios that

resemble their respective hosts' mitochondria (Table S7). However, a striking exception is the *Plasmopara viticola associated mitovirus 56*, which uses UGA/UGG in a fifty-fifty basis, while the host's mitochondria does not use UGA(Trp) at all and, in addition, shows a distinct AT content when compared to the host mitochondrial genome (57.53 versus 76.29%, respectively). A tempting hypothesis is that this virus has been originated from a fungal host that does use UGA(Trp) in relatively high levels, then switched to *Plasmopara viticola* and is still fitting its genome composition and codon usage to the new host. However, such an adaptation process would require protein synthesis to occur under a rare use of UGA(Trp) codons in the host, an aspect that should be better elucidated in the future.

Mitochondrial genomes are often biased toward a preferential use of UGA rather than UGG to encode tryptophan. This may be explained by the fact that organellar genomes, including dipteran mitochondria, usually present high AT content [94], which could in turn be the consequence of high selective pressures. Thus, synonymous codons with an A or T in the third position would mainly be selected over codons presenting C or G. Conversely, extrachromosomal genomes located in the cytoplasm would not use UGA codons for tryptophan, since they would be interpreted as stop codons by the cytoplasmic translation machinery, causing premature termination of protein synthesis. Viruses located in the mitochondria for long periods of time would progressively reflect the AT selective pressure and show increasing use of UGA over UGG. Conversely, mitoviruses that switched to a new host recently are still fitting their codon usage, presenting lower ratios of UGA/UGG. Interestingly, we did not observe a good correlation between phylogenetic relationships based on the RdRp protein and the proportion of UGA/UGG codons used for tryptophan. Thus, *Plasmopara viticola associated mitovirus 56*, Lul-MV-1, *Wenling narna-like virus 9* and *Shahe narna-like virus* are relatively close to each other but show very discrepant usage rates for UGA/UGG codons.

The fact that a viral genome presents a codon usage that resembles that of the mitochondria is a strong indication that the virus has evolved as to fit the organelle codon usage and, therefore, to use its protein synthesis machinery with high efficiency. On the other hand, dinucleotide frequency and overall AT content seem to be shaped by selective pressures that occur in the site of replication. Thus, viruses that are located and replicate within mitochondria would be more subject to compositional biases imposed in the organelle. Conversely, viruses able to colonize and replicate in the cytosol, would be less affected by selective compositional pressures, which could explain why narnaviruses and ourmiaviruses differ so much from their hosts' mitochondria not only in terms of codon usage, but also in AT content. To conclude, although an AT content bias can certainly influence the codon usage, both parameters are not totally interdependent and are probably shaped by distinct evolutionary pressures.

Another important issue concerns the virus–host interaction. An RNA virus infecting an insect is exposed to the RNA interference pathways of this host. Virus-derived small RNAs (vsRNAs) have been used in many studies as an evidence of viral infection in an organism, since they are produced through recognition of dsRNA molecules produced during the viral replication cycle [25,27,34]. Additionally, vsRNAs provide information about molecular characteristics unique for each virus species. Information based on sRNA profile such as base enrichment, size, and polarity can be used to infer the origin of the putative virus [27]. Based on the sRNA profile, we confirmed that Lul-MV-1 is replicating in the host and is not an integrated endogenous viral element (EVE), since EVE-derived small RNAs only display molecular characteristics consistent with piRNAs [26,27]. Moreover, Lul-MV-1 shows small RNA size profiles and 5′ base preference that are distinct from those observed in mitoviruses infecting fungal hosts (Figure 6A,B), but size distribution resembles that of vsRNAs from VSV, the *Vesicular stomatitis virus* that infects *Lu. longipalpis* (Figure 6C). Nevertheless, a possible phenotypic effect of narnaviruses and mitoviruses in insect hosts, as observed for some mycoviruses in fungi, is still unknown.

An important question that arises from our finding is whether Lul-MV-1 is persistent in *Lu. longipalpis* populations. Primers designed to a segment of the RdRp gene were able

to positively detect seven out of eight sand fly pools, indicating that this virus is persistent in the population (Figure 7A). Recent studies have suggested that some narnaviruses may be infective in arthropod cells, once in the tested samples, the viral RNA was greater than 0.1 per cent of total non-ribosomal RNA reads, indicating a high amount of RNA to be just a contaminant virus [15,95]. Also, the *Culex narnavirus 1* was found in different cultures of *Culex tarsalis* and sRNAs presented a peak at 21 nt in both strands, an indicative of active infection in the insect [89].

It still a matter of speculation whether mitoviruses found in arthropods are infecting the mitochondria of a fungus or protist belonging to the arthropod microbiota or the mitochondria of the arthropod itself. We believe that Lul-MV-1 is a mitovirus that infects phlebotomine mitochondria based on a body of evidence: (1) sequence similarity and close phylogenetic relationship to mitochondrial viruses of the genus *Mitovirus*; (2) the viral genome presents a high AT content (69.74%), similarly to what is observed in organellar genomes, including the mtDNA of *Lu. longipalpis* (78.07%); (3) codon usage is closer to mitochondria of invertebrate hosts than to fungal hosts; (4) the virus uses mainly the codon UGA to code tryptophan, in consonance with the codon usage of of *Lu. longipalpis* mitochondria; (5) dinucleotide composition of Lul-MV-1 genome resembles the composition of insect genomes rather than fungal genomes; (6) virus-derived sRNAs suggests activation of siRNA pathway in insects rather than fungal hosts; (7) RT-PCR followed by Sanger sequencing, confirmed the presence of the viral genome and the prevalence of Lul-MV-1 in a sand fly laboratory colony; and lastly (8) reads from different RNA and DNA public libraries of *Lu. longipalpis* were successfully mapped on the viral genome, confirming the high prevalence of the virus in *Lu. longipalpis* populations.

Altogether, the experimental and bioinformatic methods applied in this study allowed us to detect, classify, and characterize a novel mitovirus infecting the mitochondria of the sand fly *Lu. longipalpis*. In addition to compositional and phylogenetic analyses, the utilization of vsRNA profiles represent a valuable approach to properly ascribe the respective hosts of viruses detected in metagenomic datasets. According to our knowledge, this is the first report of a mitovirus infecting an insect host, and the results presented herein highlight the large diversity of the virosphere and the possibility that mitoviruses may infect a much wider range of hosts than initially supposed.

References

1. Hyman, P.; Abedon, S.T. Smaller Fleas: Viruses of Microorganisms. *Scientifica* **2012**, *2012*, 1–23. [CrossRef]
2. Zhang, Y.-Z.; Shi, M.; Holmes, E.C. Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **2018**, *172*, 1168–1172. [CrossRef]
3. Graham, E.B.; Paez-Espino, D.; Brislawn, C.; Hofmockel, K.S.; Wu, R.; Kyrpides, N.C.; Jansson, J.K.; McDermott, J.E. Untapped viral diversity in global soil metagenomes. *bioRxiv* 2019. [CrossRef]
4. Ng, T.F.F.; Willner, D.L.; Lim, Y.W.; Schmieder, R.; Chau, B.; Nilsson, C.; Anthony, S.; Ruan, Y.; Rohwer, F.; Breitbart, M. Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes. *PLoS ONE* **2011**, *6*, e20579. [CrossRef]
5. Lim, E.S.; Zhou, Y.; Zhao, G.; Bauer, I.K.; Droit, L.; Ndao, I.M.; Warner, B.B.; Tarr, P.I.; Wang, D.; Holtz, L.R. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **2015**, *21*, 1228–1234. [CrossRef]
6. Reyes, A.P.; Alves, J.M.; Durham, A.M.; Gruber, A. Use of profile hidden Markov models in viral discovery: Current insights. *Adv. Genom Genet.* **2017**, *7*, 29–45. [CrossRef]
7. Holland, J.; Spindler, K.; Horodyski, F.; Grabau, E.; Nichol, S.; VandePol, S. Rapid evolution of RNA genomes. *Science* **1982**, *215*, 1577–1585. [CrossRef]
8. Drake, J.W. Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 4171–4175. [CrossRef] [PubMed]
9. Peck, K.M.; Lauring, A.S. Complexities of Viral Mutation Rates. *J. Virol* **2018**, *92*, e01031-17. [CrossRef]
10. Sanjuán, R.; Nebot, M.R.; Chirico, N.; Mansky, L.M.; Belshaw, R. Viral Mutation Rates. *JVI* **2010**, *84*, 9733–9748. [CrossRef]
11. Fancello, L.; Raoult, D.; Desnues, C. Computational tools for viral metagenomics and their application in clinical research. *Virology* **2012**, *434*, 162–174. [CrossRef]
12. Handelsman, J.; Rondon, M.R.; Brady, S.F.; Clardy, J.; Goodman, R.M. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* **1998**, *5*, R245–R249. [CrossRef]
13. Chiu, C.Y. Viral pathogen discovery. *Curr. Opin. Microbiol.* **2013**, *16*, 468–478. [CrossRef]
14. Pallen, M.J. Diagnostic metagenomics: Potential applications to bacterial, viral and parasitic infections. *Parasitology* **2014**, *141*, 1856–1862. [CrossRef]
15. Shi, M.; Lin, X.-D.; Tian, J.-H.; Chen, L.-J.; Chen, X.; Li, C.-X.; Qin, X.-C.; Li, J.; Cao, J.-P.; Eden, J.-S.; et al. Redefining the invertebrate RNA virosphere. *Nature* **2016**, *540*, 539–543. [CrossRef]
16. Roux, S.; Chan, L.-K.; Egan, R.; Malmstrom, R.R.; McMahon, K.D.; Sullivan, M.B. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **2017**, *8*, 858. [CrossRef]
17. Novella, I.S.; Clarke, D.K.; Quer, J.; Duarte, E.A.; Lee, C.H.; Weaver, S.C.; Elena, S.F.; Moya, A.; Domingo, E.; Holland, J.J. Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. *J. Virol.* **1995**, *69*, 6805–6809. [CrossRef]
18. Belalov, I.S.; Lukashev, A.N. Causes and Implications of Codon Usage Bias in RNA Viruses. *PLoS ONE* **2013**, *8*, e56642. [CrossRef]
19. Lobo, F.P.; Mota, B.E.F.; Pena, S.D.J.; Azevedo, V.; Macedo, A.M.; Tauch, A.; Machado, C.R.; Franco, G.R. Virus–host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts. *PLoS ONE* **2009**, *4*, e6282. [CrossRef]
20. Biswas, K.; Palchoudhury, S.; Chakraborty, P.; Bhattacharyya, U.; Ghosh, D.; Debnath, P.; Ramadugu, C.; Keremane, M.; Khetarpal, R.; Lee, R. Codon Usage Bias Analysis of Citrus tristeza Virus: Higher Codon Adaptation to Citrus reticulata Host. *Viruses* **2019**, *11*, 331. [CrossRef]
21. Chen, M.; Tan, Z.; Zeng, G.; Peng, J. Comprehensive Analysis of Simple Sequence Repeats in Pre-miRNAs. *Mol. Biol. Evol.* **2010**, *27*, 2227–2232. [CrossRef]
22. Chen, M.; Tan, Z.; Zeng, G. Microsatellite is an important component of complete Hepatitis C virus genomes. *Infect. Genet. Evol.* **2011**, *11*, 1646–1654. [CrossRef]
23. Di Giallonardo, F.; Schlub, T.E.; Shi, M.; Holmes, E.C. Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *J. Virol* **2017**, *91*, e02381-16. [CrossRef]
24. Obbard, D.J.; Gordon, K.H.J.; Buck, A.H.; Jiggins, F.M. The evolution of RNAi as a defence against viruses and transposable elements. *Phil. Trans. R. Soc. B* **2009**, *364*, 99–115. [CrossRef]

25. Aguiar, E.R.G.R.; Olmo, R.P.; Paro, S.; Ferreira, F.V.; de Faria, I.J.D.S.; Todjro, Y.M.H.; Lobo, F.P.; Kroon, E.G.; Meignin, C.; Gatherer, D.; et al. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic Acids Res.* **2015**, *43*, 6191–6206. [CrossRef]

26. Aguiar, E.R.G.R.; de Almeida, J.P.P.; Queiroz, L.R.; Oliveira, L.S.; Olmo, R.P.; de Faria, I.J.; da S Imler, J.-L.; Gruber, A.; Matthews, B.J.; Marques, J.T. A single unidirectional piRNA cluster similar to the flamenco locus is the major source of EVE-derived transcription and small RNAs in Aedes aegypti mosquitoes. *RNA* **2020**, *26*, 581–594. [CrossRef]

27. Aguiar, E.R.G.R.; Olmo, R.P.; Marques, J.T. Virus-derived small RNAs: Molecular footprints of host-pathogen interactions: Virus-derived small RNAs. *WIREs RNA* **2016**, *7*, 824–837. [CrossRef]

28. Webster, C.L.; Waldron, F.M.; Robertson, S.; Crowson, D.; Ferrari, G.; Quintana, J.F.; Brouqui, J.-M.; Bayne, E.H.; Longdon, B.; Buck, A.H.; et al. The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. *PLoS Biol* **2015**, *13*, e1002210. [CrossRef]

29. Fukuda, M.M.; Klein, T.A.; Kochel, T.; Quandelacy, T.M.; Smith, B.L.; Villinski, J.; Bethell, D.; Tyner, S.; Se, Y.; Lon, C.; et al. Malaria and other vector-borne infection surveillance in the U.S. Department of Defense Armed Forces Health Surveillance Center-Global Emerging Infections Surveillance program: Review of 2009 accomplishments. *BMC Public Health* **2011**, *11*, S9. [CrossRef] [PubMed]

30. Gould, E.; Pettersson, J.; Higgs, S.; Charrel, R.; de Lamballerie, X. Emerging arboviruses: Why today? *One Health* **2017**, *4*, 1–13. [CrossRef]

31. Ayhan, N.; Prudhomme, J.; Laroche, L.; Bañuls, A.-L.; Charrel, R.N. Broader Geographical Distribution of Toscana Virus in the Mediterranean Region Suggests the Existence of Larger Varieties of Sand Fly Vectors. *Microorganisms* **2020**, *8*, 114. [CrossRef]

32. Minard, G.; Mavingui, P.; Moro, C. Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasit Vectors* **2013**, *6*, 146. [CrossRef]

33. Calisher, C.H.; Higgs, S. The Discovery of Arthropod-Specific Viruses in Hematophagous Arthropods: An Open Door to Understanding the Mechanisms of Arbovirus and Arthropod Evolution? *Annu. Rev. Entomol.* **2018**, *63*, 87–103. [CrossRef] [PubMed]

34. Ferreira, F.V.; Aguiar, E.R.G.R.; Olmo, R.P.; de Oliveira, K.P.V.; Silva, E.G.; Sant'Anna, M.R.V.; Gontijo, N.D.F.; Kroon, E.G.; Imler, J.L.; Marques, J.T. The small non-coding RNA response to virus infection in the Leishmania vector Lutzomyia longipalpis. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0006569. [CrossRef] [PubMed]

35. Cook, S.; Chung, B.Y.-W.; Bass, D.; Moureau, G.; Tang, S.; McAlister, E.; Culverwell, C.L.; Glücksman, E.; Wang, H.; Brown, T.D.K.; et al. Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts. *PLoS ONE* **2013**, *8*, e80720. [CrossRef] [PubMed]

36. Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses. In *International Committee on Taxonomy of Viruses*; King, A.M.Q. (Ed.) Academic Press: Waltham, MA, USA, 2012; ISBN 978-0-12-384684-6.

37. Hillman, B.I.; Cai, G. The Family Narnaviridae. In *Advances in Virus Research*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 86, pp. 149–176. ISBN 978-0-12-394315-6.

38. ICTV Master Species List 2019.v1 (MSL #35). Available online: https://talk.ictvonline.org/files/master-species-lists/m/msl/9601/download (accessed on 27 August 2020).

39. Koonin, E.V.; Dolja, V.V. Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements. *Microbiol. Mol. Biol. Rev.* **2014**, *78*, 278–303. [CrossRef]

40. Mizutani, Y.; Abraham, A.; Uesaka, K.; Kondo, H.; Suga, H.; Suzuki, N.; Chiba, S. Novel Mitoviruses and a Unique Tymo-Like Virus in Hypovirulent and Virulent Strains of the Fusarium Head Blight Fungus, Fusarium boothii. *Viruses* **2018**, *10*, 584. [CrossRef] [PubMed]

41. Nibert, M.; Debat, H.; Manny, A.; Grigoriev, I.; De Fine Licht, H. Mitovirus and Mitochondrial Coding Sequences from Basal Fungus Entomophthora muscae. *Viruses* **2019**, *11*, 351. [CrossRef]

42. Espino-Vázquez, A.N.; Bermúdez-Barrientos, J.R.; Cabrera-Rangel, J.F.; Córdova-López, G.; Cardoso-Martínez, F.; Martínez-Vázquez, A.; Camarena-Pozos, D.A.; Mondo, S.J.; Pawlowska, T.E.; Abreu-Goodger, C.; et al. Narnaviruses: Novel players in fungal–bacterial symbioses. *ISME J.* **2020**, *14*, 1743–1754. [CrossRef]

43. Lin, Y.; Zhou, J.; Zhou, X.; Shuai, S.; Zhou, R.; An, H.; Fang, S.; Zhang, S.; Deng, Q. A novel narnavirus from the plant-pathogenic fungus Magnaporthe oryzae. *Arch. Virol.* **2020**, *165*, 1235–1240. [CrossRef]

44. Pearson, M.N.; Beever, R.E.; Boine, B.; Arthur, K. Mycoviruses of filamentous fungi and their relevance to plant pathology. *Mol. Plant. Pathol.* **2009**, *10*, 115–128. [CrossRef] [PubMed]

45. Dolja, V.V.; Koonin, E.V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **2018**, *244*, 36–52. [CrossRef] [PubMed]

46. Ohkita, S.; Lee, Y.; Nguyen, Q.; Ikeda, K.; Suzuki, N.; Nakayashiki, H. Three ourmia-like viruses and their associated RNAs in Pyricularia oryzae. *Virology* **2019**, *534*, 25–35. [CrossRef] [PubMed]

47. Mokili, J.L.; Rohwer, F.; Dutilh, B.E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2012**, *2*, 63–77. [CrossRef]

48. Brenner, S.E.; Chothia, C.; Hubbard, T.J.P. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6073–6078. [CrossRef]

49. Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; Chothia, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **1998**, *284*, 1201–1210. [CrossRef]
50. Skewes-Cox, P.; Sharpton, T.J.; Pollard, K.S.; DeRisi, J.L. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. *PLoS ONE* **2014**, *9*, e105067. [CrossRef]
51. Van der Auwera, S.; Bulla, I.; Ziller, M.; Pohlmann, A.; Harder, T.; Stanke, M. ClassyFlu: Classification of Influenza A Viruses with Discriminatively Trained Profile-HMMs. *PLoS ONE* **2014**, *9*, e84558. [CrossRef]
52. Masembe, C.; Phan, M.V.T.; Robertson, D.L.; Cotten, M. Increased resolution of African swine fever virus genome patterns based on profile HMMs of protein domains. *Virus Evol.* **2020**, *6*, veaa044. [CrossRef]
53. Bramley, J.C.; Yenkin, A.L.; Zaydman, M.A.; DiAntonio, A.; Milbrandt, J.D.; Buchser, W.J. Domain-centric database to uncover structure of minimally characterized viral genomes. *Sci. Data* **2020**, *7*, 202. [CrossRef]
54. Alves, J.M.P.; de Oliveira, A.L.; Sandberg, T.O.M.; Moreno-Gallego, J.L.; de Toledo, M.A.F.; de Moura, E.M.M.; Oliveira, L.S.; Durham, A.M.; Mehnert, D.U.; de A. Zanotto, P.M.; et al. GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. *Front. Microbiol.* **2016**, *7*. [CrossRef] [PubMed]
55. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [CrossRef] [PubMed]
56. Altschul, S. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
57. Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J.; McQuillan, J.A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **2012**, *28*, 464–469. [CrossRef]
58. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]
59. Ladner, J.T.; Beitzel, B.; Chain, P.S.G.; Davenport, M.G.; Donaldson, E.; Frieman, M.; Kugelman, J.; Kuhn, J.H.; O'Rear, J.; Sabeti, P.C.; et al. Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing. *mBio* **2014**, *5*, e01360-14. [CrossRef]
60. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
61. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
62. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef]
63. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [CrossRef]
64. Huson, D.H.; Scornavacca, C. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Syst. Biol.* **2012**, *61*, 1061–1067. [CrossRef] [PubMed]
65. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]
66. Marques, J.T.; Wang, J.-P.; Wang, X.; de Oliveira, K.P.V.; Gao, C.; Aguiar, E.R.G.R.; Jafari, N.; Carthew, R.W. Functional Specialization of the Small Interfering RNA Pathway in Response to Virus Infection. *PLoS Pathog.* **2013**, *9*, e1003579. [CrossRef]
67. Wei, T.; Simko, V. Package 'corrplot'. Available online: https://cran.r-project.org/web/packages/corrplot/corrplot.pdf (accessed on 15 August 2020).
68. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
69. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Use R! Springer: New York, NY, USA, 2009; ISBN 978-0-387-98140-6.
70. Maumus, F.; Fiston-Lavier, A.-S.; Quesneville, H. Impact of transposable elements on insect genomes and biology. *Curr. Opin. Insect Sci.* **2015**, *7*, 30–36. [CrossRef]
71. Aiewsakun, P.; Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **2015**, *479–480*, 26–37. [CrossRef]
72. Ballinger, M.J.; Bruenn, J.A.; Taylor, D.J. Phylogeny, integration and expression of sigma virus-like genes in Drosophila. *Mol. Phylogenet. Evol.* **2012**, *65*, 251–258. [CrossRef]
73. Virus Taxonomy: The Classification and Nomenclature of Viruses The 9th Report of the ICTV (2011). Available online: https://talk.ictvonline.org/ictv-reports/ictv_9th_report/ (accessed on 27 August 2020).
74. Shackelton, L.A.; Holmes, E.C. The role of alternative genetic codes in viral evolution and emergence. *J. Theor. Biol.* **2008**, *254*, 128–134. [CrossRef]
75. Bruenn, J.A.; Warner, B.E.; Yerramsetty, P. Widespread mitovirus sequences in plant genomes. *PeerJ* **2015**, *3*, e876. [CrossRef]
76. Chiapello, M.; Rodríguez-Romero, J.; Ayllón, M.A.; Turina, M. Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evol.* **2020**, veaa058. [CrossRef]
77. Zoll, J.; Verweij, P.E.; Melchers, W.J.G. Discovery and characterization of novel Aspergillus fumigatus mycoviruses. *PLoS ONE* **2018**, *13*, e0200511. [CrossRef] [PubMed]

78. DeRisi, J.L.; Huber, G.; Kistler, A.; Retallack, H.; Wilkinson, M.; Yllanes, D. An exploration of ambigrammatic sequences in narnaviruses. *Sci. Rep.* **2019**, *9*, 17982. [CrossRef] [PubMed]

79. Dinan, A.M.; Lukhovitskaya, N.I.; Olendraite, I.; Firth, A.E. A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol.* 2020. [CrossRef]

80. Velazquez-Salinas, L.; Zarate, S.; Eschbaumer, M.; Pereira Lobo, F.; Gladue, D.P.; Arzt, J.; Novella, I.S.; Rodriguez, L.L. Selective Factors Associated with the Evolution of Codon Usage in Natural Populations of Arboviruses. *PLoS ONE* **2016**, *11*, e0159943. [CrossRef] [PubMed]

81. Schuster, S.; Miesen, P.; van Rij, R.P. Antiviral RNAi in Insects and Mammals: Parallels and Differences. *Viruses* **2019**, *11*, 448. [CrossRef] [PubMed]

82. Ayhan, N.; Charrel, R.N. Of phlebotomines (sandflies) and viruses: A comprehensive perspective on a complex situation. *Curr. Opin. Insect Sci.* **2017**, *22*, 117–124. [CrossRef]

83. Alkan, C.; Bichaud, L.; de Lamballerie, X.; Alten, B.; Gould, E.A.; Charrel, R.N. Sandfly-borne phleboviruses of Eurasia and Africa: Epidemiology, genetic diversity, geographic range, control measures. *Antivir. Res.* **2013**, *100*, 54–74. [CrossRef]

84. Nibert, M.L.; Vong, M.; Fugate, K.K.; Debat, H.J. Evidence for contemporary plant mitoviruses. *Virology* **2018**, *518*, 14–24. [CrossRef]

85. Stough, J.M.A.; Beaudoin, A.J.; Schloss, P.D. Coding-Complete RNA Virus Genomes Assembled from Murine Cecal Metatran-scriptomes. *Microbiol. Resour. Announc.* **2020**, *9*, e00018-20. [CrossRef]

86. Richaud, A.; Frézal, L.; Tahan, S.; Jiang, H.; Blatter, J.A.; Zhao, G.; Kaur, T.; Wang, D.; Félix, M.-A. Vertical transmission in Caenorhabditis nematodes of RNA molecules encoding a viral RNA-dependent RNA polymerase. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 24738–24747. [CrossRef]

87. Mahar, J.E.; Shi, M.; Hall, R.N.; Strive, T.; Holmes, E.C. Comparative Analysis of RNA Virome Composition in Rabbits and Associated Ectoparasites. *J. Virol.* **2020**, *94*, e02119-19. [CrossRef] [PubMed]

88. Chandler, J.A.; Liu, R.M.; Bennett, S.N. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. *Front. Microbiol.* **2015**, *6*. [CrossRef] [PubMed]

89. Göertz, G.; Miesen, P.; Overheul, G.; van Rij, R.; van Oers, M.; Pijlman, G. Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in Aedes and Culex Mosquito Cells. *Viruses* **2019**, *11*, 271. [CrossRef] [PubMed]

90. Nibert, M.L. Mitovirus UGA (Trp) codon usage parallels that of host mitochondria. *Virology* **2017**, *507*, 96–100. [CrossRef] [PubMed]

91. Je, M.; Kim, H.; Son, H.S. Analysis of the codon usage pattern of the RdRP gene of mycovirus infecting *Aspergillus* spp. *Virol. J.* **2019**, *16*, 10. [CrossRef] [PubMed]

92. Chen, F.; Wu, P.; Deng, S.; Zhang, H.; Hou, Y.; Hu, Z.; Zhang, J.; Chen, X.; Yang, J.-R. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. *Nat. Ecol. Evol.* **2020**, *4*, 589–600. [CrossRef] [PubMed]

93. Nie, Y.; Wang, L.; Cai, Y.; Tao, W.; Zhang, Y.-J.; Huang, B. Mitochondrial genome of the entomophthoroid fungus Conidiobolus heterosporus provides insights into evolution of basal fungi. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 1379–1391. [CrossRef] [PubMed]

94. Ramakodi, M.P.; Singh, B.; Wells, J.D.; Guerrero, F.; Ray, D.A. A 454 sequencing approach to dipteran mitochondrial genome research. *Genomics* **2015**, *105*, 53–60. [CrossRef]

95. Shi, M.; Neville, P.; Nicholson, J.; Eden, J.-S.; Imrie, A.; Holmes, E.C. High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J. Virol.* **2017**, *91*, e00680-17. [CrossRef]