bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

New Results

Article Category: Microbiology

An RNA Virome associated to the Golden orb-weaver Spider *Nephila clavipes*

Authors:

Humberto J. Debat[1*]

Affiliations:

[1] *Instituto de Patología Vegetal, Centro de Investigaciones Agropecuarias, Instituto Nacional de Tecnología Agropecuaria (IPAVE-CIAP-INTA), X5020ICA, Córdoba, Argentina*

* *Corresponding author.* At: IPAVE-CIAP-INTA, 11 de setiembre 4755, X5020ICA, Córdoba, Argentina.

Email address: debat.humberto@inta.gob.ar

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The golden orb-weaver *Nephila clavipes* is an abundant and widespread, sexual dimorphic spider species. The first annotated genome of orb-weaver spiders, exploring *N. clavipes,* has been reported recently. This remarkable study, focused primarily in the diversity of silk specific genes, shed light into the complex evolutionary history of spiders. Furthermore, a robust, multiple and tissue specific transcriptome analysis provided a massive resource for *N. clavipes* RNA survey. Here, I present evidence towards the discovery and characterization of viral sequences corresponding to the first extant virus species associated to *N. clavipes* and the *Nephilidae* family. The putative new species are linked to ssRNA positive-strand viruses, such as Picornavirales, but also to ssRNA negative-strand and dsRNA viruses. In addition, I detected sequence data of new strains of two recently reported arthropod viruses, which complemented and extended the corresponding sequence references. The identified viruses appear to be complete, potentially functional, and presenting the typical architecture and consistent viral domains. The intrinsic nature of the detected sequences and their absence in the recently generated genome assembly, suggest that they correspond to *bona fide* RNA virus sequences. The available RNA data allowed for the first time to address a tissue/organ specific analysis of virus loads/presence in spiders, suggesting a complex spatial and differential distribution of the tentative viruses, encompassing the spider brain, and also silk and venom glands. Until recently, the virus landscape associated to spiders remained elusive. The discovered viruses described here, provide only a fragmented glimpse of the potential magnitude of the *Aranea* virosphere. Future studies should focus not only on complementing and expanding these preceding findings, but also on addressing the potential ecological role of these viruses, which might influence the biology of these outstanding arthropod species.

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The golden orb-weaver *Nephila clavipes* (Linnaeus, 1767) is a female-biased sexual-size dimorphic spider (Kuntner, 2009). It is a widespread and abundant species, distributed from southeastern United States to northern Argentina and from the Galapagos Islands to the Caribbean. They inhabit a broad range of habitats that vary from mild to strong seasonality (Higgins, 2000). *N. clavipes* spiders use golden-colored silks to spin orb webs. They are opportunistic predators, capturing diverse arthropods, and even small vertebrates (Higgins et al., 1992). Spider silks have a great potential for medical and industrial innovation, given their features of being both extremely strong and light (Agnarsson et al., 2010). *N. calvipes* generates a battery of silks derived from seven types of araneoid silk glands, thus is considered the "ubiquitous workhorse of silk spider research" (Vollrath, 2000). Despite the importance of this extensively studied spider, the molecular characterization of its genetic repertoire was lacking in the literature. Babb et al (2017) recently reported a sequencing *tour de force* that generated the first annotated genome of *N. calvipes*. Besides generating a genome assembly of 2.8 GB, the authors explored the RNA component of *N. clavipes* by multiple RNA-seq data from 16 different tissue/organ/individual isolates (whole body, brain, and silk and venom glands) collected from four female individuals. An integrated *de novo* assembled transcriptome was generated for each isolate using strand-specific, ribosomal RNA (rRNA)-depleted, 100-bp paired-end reads. In sum, a total of 1,848,260,474 raw RNA reads were quality controlled and filtered, and the curated 1,531,402,748 reads were *de novo* assembled using Trinity (rel_2.25.13) yielding 1,507,505 unique strand-specific transcripts. These 1.53×10^9 reads corresponding to an assembly of 1.5 million transcripts, which represent the most extensive deposited RNA assembled data of any organism at the NCBI TSA database, were used as input for the objective of this study: The identification and characterization of potential RNA viruses associated to *N. clavipes*, which remain elusive from the literature.

Results and Discussion

In order to identify putative RNA virus associated with *N. clavipes*, the complete RefSeq release of viral sequences was retrieved from ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/. The *N. clavipes* 1.5 million transcripts RNA assembly was assessed by multiple TBLASTN searches (max e-value = 1×10^{-5}) using as probe the complete predicted non redundant viral Refseq proteins in a local server. Significant hits were massively explored by hand, and redundant contigs

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(3,547) discarded. Potential open reading frames (ORF) were predicted by ORFfinder, translated putative proteins were blasted against the non-redundant protein sequences NR database, and best hits were retrieved. Based on sequence homology to best hits, sequence alignments, predicted proteins and domains, and phylogenetic comparisons to reported species, I found evidence of 6 diverse new virus species and 3 new strains of reported virus species associated to *N. clavipes* (Figure 1; Supp. Figure 1).

Nephile clavipes virus 1

The proposed *Nephile clavipes* virus 1 (NcV1) is predicted to have a 10,198 nt long genome, presenting a single ORF between coordinates 761-10,147, encoding a putative 357.963 kDa and 3,128 aa long polyprotein (Supp. Figure 2). Domain prediction based on InterproScan, NCBI-CD database v3.15, THHM, PHOBIUS, SMART, Pfam, PROSITE and Garnier resulted in the identification of diverse motifs associated to ssRNA positive strand viruses, most specifically from the Picornavirales order (Le Gall et al., 2008). For instance, between the 688-861 aa and 1,120-1,342 aa coordinates a Calici_coat (Pfam id PF00915) and a CRPV_capsid (Pfam id PF08762) were predicted, associated to the coat protein of *Caliciviridae* family and capsid protein of *Dicistroviridae* family respectively, the latter being a member assigned to the Picornavirales order. Three additional domains were found between the 1,722-1,886 aa , 2,356-2,553 aa and 2,614-3,094 aa coordinates corresponding to a viral RNA helicase (Pfam id PF00910), a Trypsin-like serine protease (InterPro id IPR009003) and a RNA dependent RNA Polymerase (RdRP; RdRP_1 Pfam id PF00680) found in many positive strand RNA eukaryotic viruses. In addition, by assessing sequence similarity I suggest that NcV1 is more closely related to Wuhan spider virus 2 (WSV2) (Pairwise % Identity: 56.6% at the RNA level, and 39.5% at the predicted polyprotein level). WSV2 is currently unclassified, but tentatively assigned to the Picornavirales, within a newly proposed super clade of Picorna-Calici by the reporting authors (Shi et al., 2016). It is important to highlight that WSV2 was identified recently from a RNA pooled sample, a sequencing library conformed by individuals of diverse classified and unclassified spiders: *Neoscona nautica* (14), *Parasteatoda tepidariorum* (3), *Plexippus setipes* (3), *Pirata sp.* (1), and 8 unrecognized individuals (*Araneae sp.*) (Shi et al., 2016). Thus, the specific host of WSV2 remains to be determined, even at the family level, which could contribute to understanding the evolutionary history of these viruses and their hosts.

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nephile clavipes virus 2

Nephile clavipes virus 2 (NcV2) shares the proposed Picorna-Calici ssRNA(+) super clade with NcV1, but presents a different genome organization and domain architecture (Supp. Figure 3). NcV2 presents an 11,622 nt RNA genome enclosing 4 putative ORFs. ORF1 (722-7,226 nt coordinates) encodes a putative replicase protein (RP) of 274 kDa and 2,412 aa long. Several domains were found along the RP, between the 172-443 aa, 927-1,047 aa, 1,697-1,841 aa and 1,901-2,400 aa coordinates, corresponding to a viral RNA helicase (Pfam id PF00910), a P-loop containing nucleoside triphosphate hydrolases (InterPro id IPR027417), a Trypsin-like serine protease (InterPro id IPR009003) and a RdRP (RdRP_1 Pfam id PF00680) respectively. NcV2 ORF2 (7,948-9,747 nt) encodes a putative 67.3 kDa coat protein, 599 aa long, presenting a Calici_coat domain (Pfam id PF00915) and a picornavirus capsid protein domain Rhv (Pfam id pfam00073) associated to Picornavirales. ORF3 (9,792-10,562 nt) encodes a 27.7 kDa, 256 aa long protein similar to hypothetical protein 3 of Hubei picorna-like virus 76 (E-value = 8e-14, 38% identity) and of Wuhan spider virus 6 (WSV6; E-value = 8e-14, 38% identity), both of unknown function. ORF 4 (10,702-11,457 nt) encodes a 29.7 kDa, 251 aa long protein, with a central coiled-coil region, and of unknown function. Although is evident that NcV1 and NcV2 share several conserved domains, that could suggest a common origin, their divergent spatial arrangement and their low sequence identity might be interpreted as a timely ancestral separation followed by recombination and reorganization of the putative proteins on the respective viruses. NcV2 is more closely related to WSV6 (Pairwise Identity: 51.7% at the RNA level, and 28.5% at the predicted replicase protein). WSV6 was also identified recently from a pooled sample library conformed by individuals of diverse spiders, has been provisionally assigned to a Picorna-calici superclade, and its specific host spider remains unidentified. Both NcV1 and NcV2 when explored by maximum likelihood phylogenetic trees derived from MAFFT alignments of replicase proteins, tend to cluster among several newly found picorna-like invertebrate virus. The related literature is exceptionally limited, thus until future studies explore the potential impact of these viruses on their host, it might be prudent not to speculate on their biology. Associated literature is available only regarding slightly related virus species of this order which have been studied in more detail, such as *Acute bee paralysis virus* (*Picornavirales; Dicistroviridae; Aparavirus*) or *Sacbrood Virus* (*Picornavirales; Iflaviridae; Iflavirus)* which are reported to have drastic effect on their bee host, resulting in larvae death and sudden colony collapse (Govan et al., 2000; Ghosh et al., 1999).

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nephile clavipes virus 3

The putative *Nephila clavipes* virus 3 (NcV3) is predicted to have an 8,138 nt long ssRNA (+) strand genome, presenting 2 partially overlapping ORFs. ORF1 (9-6,809 nt coordinates) encodes a putative replicase protein (RP) of 260.4 kDa and 2,266 aa long (Supp. Figure 4). Several domains were found along the RP, between the 131-521 aa, 1,233-1,573 aa, and 1,786-2,247 aa coordinates corresponding to an alpha-virus like viral methyl-transferasa (Pfam id PF01660), a (+) RNAvirus RNA helicase (Pfam id PF01443), and a Tymovirus-like RdRP (RdRP_2 Pfam id PF00978) respectively. ORF2 (6,736-8,088 nt coordinates), which is tentatively translated by ribosomal frameshifting, encodes a 51.8 kDa, 450 aa long hypothetical protein, sharing significant identity with the virion structural glycoprotein s2gp2 (E-value = 3e-05, 41% identity) corresponding to RNA 2 of the bisegmented *Chronic bee paralysis virus* (CBPV). To date, the ssRNA (+) CBPV remains unclassified by the International Committee on Taxonomy of Viruses (ICTV), and only its RP sequence presents similarities with members of the *Nodaviridae* and *Tombusviridae* families. It has been suggested that CBPV might be the prototype species of a new family of positive single-stranded RNA viruses (Olivier et al., 2008). MAFFT alignments of the replicase protein of NcV3, followed with FastTree maximum likelihood phylogenetic trees suggest that NcV3 is further related to *Virgaviridae* distant like viruses, such as the proposed Lodeiro virus (LV) and Hubei virga-like virus 9. LV is an unclassified virus that has been recently reported to be derived from the crab spider *Philodromus dispar* (*Philodromidae*) (Shean et al., 2017). NcV3 shares with LV a 51.8% nt sequence identity at the RNA level, and 28.4% at the RP protein. The LV genome appears to be 1.7 kb longer than NcV3, presenting three additional smalls ORFs that seem to be absent in NcV3. The only additional LV ORF with a predicted function (ORF4) is predicted to be a virus RNA helicase. NcV3 presents a related RNA helicase domain with a P-loop NTPase region (InterPro id IPR027417) at the central region of the RP, which could replace the function associated to the predicted ORF4 protein of LV.

Nephile clavipes virus 4

Nephila clavipes virus 4 (NcV4) shares with NcV3 a super clade of Virga-like viruses. However the divergent NcV4 clusters within a distinct group of ssRNA(+) viruses, some of them associated with nematodes, such as the recently reported Xinzhou nematode virus 1 (XzNV1) and Xingshan nematode virus 1 (XgNV1; Shi et al., 2016). NcV4 presents an 11,919 nt ssRNA (+)

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genome enclosing three putative ORFs, of similar size and organization as XzNV1 and XgNV1. ORF1 (90-8,843 nt coordinates) encodes a putative replicase protein (RP) of 334.9 kDa and 2,917 aa long (Supp. Figure 5). Several distinct functional domains were predicted covering the RP, between the 51-429 aa, 1,141-1,267 aa, 2,045-2,305 aa and 2,454-2,894 aa coordinates, corresponding to a an alpha-virus like viral methyl-transferasa (Pfam id PF01660), a Clavaminate synthase-like domain (Superfamily db id SSF51197), a (+) RNAvirus RNA helicase (Pfam id PF01443), and a Tymovirus-like RdRP (RdRP_2 Pfam id PF00978) respectively. NcV4 ORF2 (8,934-11,051 nt) encodes a putative 80.7 kDa, 705 aa long hypothetical protein harboring transmembrane signal at the N-terminus predicted by PHOBIUS, similar to the hypothetical protein 2 of Hubei virga-like virus 13 (E-value = 2e-108, 35% identity). ORF3 (11,084-11,818 nt) encodes a 27.3 kDa, 244 aa long protein, akin to the hypothetical protein 3 of XgNV1 (E-value = 1e-20, 32% identity). Both ORF3 and ORF4 putative proteins are probably structural of unknown function.

Nephile clavipes virus 5

The putative *Nephila clavipes* virus 5 (NcV5) is predicted to have a 7,369 nt long genome, presenting a single ORF between coordinates 56-7,285 nt (3´to 5´orientation), encoding a putative 278.9 kDa and 2,409 aa long polyprotein (Supp. Figure 6). Based on sequence structure and domain prediction of the putative protein, NcV5 appears to be an ssRNA (-) virus, distantly related to the *Bunyaviridae* family (*Mononegavirales*), and NcV5 predicted protein resembles de RdRP encoded typically in the Large genome segment of this multipartite viruses. Nevertheless, NcV5 is more closely related to several new "bunya-like RNA viruses" which are suggested to have lost during evolution many genome segments, and thus harboring only an RdRP. In fact, NcV5 is similar to the proposed uni-segmented Shayang spider virus 2 (E-value = 7e-25, pairwise identity at the RdRP = 30%) (Li et al., 2015). Relaxed sequence TBLASTN searches based on G, Ns, Nm proteins of multipartite related *Bunyaviridae* failed to retrieve any other potential gene segments of NcV5. Hence, only temporarily and until further studies generate conclusive evidence, NcV5 is speculated to be a monosegmented ssRNA (-) virus. The RdRP of NcV5 presents a RNA polymerase associated to bunyavirus domain (Pfam id PF04196) at the 897-1,410 aa region, and a congruent RDRP of ssRNA(-) signature (InterPro id IPR007099). MAFFT alignments of the RdRP of NcV5, followed with FastTree maximum likelihood phylogenetic trees of related refseq sequences, suggest that NcV5 might be a member of a new clade of spider bunya-like mono-

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

segmented viruses, pivoting distantly to *Tospovirus* and slightly closer to *Orthobunyavirus* and *Hantavirus* (*Bunyaviridae*). *Orthobunyavirus* are vectored by mosquitoes and also by Ticks (*Arachnida; Acari*), thus there could be some grounds to postulate that *N. clavipes* might have acquired the potential to host a replicative form of this bunya-like NcV5.

Nephile clavipes virus 6

The proposed *Nephile clavipes* virus 6 (NcV6) is predicted to have a bi-segmented genome. RNA segment 1 is 4,006 nt long, presenting a single ORF between coordinates 10-4,005 nt, encoding a putative 152.4 kDa, 1,331 aa long RP (Supp. Figure 7). NcV6 presents only at the central region of the RP a RNA polymerase signature (Superfamily id SSF56672). NcV6 appears to be highly divergent, and only based on BLASTP searches it could be poorly associated with dsRNA viruses, specifically to RNA segment 1 of *Homalodisca vitripennis reovirus* (*Reoviridae; Sedoreovirinae; Phytoreovirus*) sharing a 23% aa similarity at the RP (E-value = 3e-26). Hidden Markov Models searches hint that the NcV6 RP is related to the Hubei reo-like virus 10 (HrlV10) RdRP (E-value = 4e-35), a recently described dsRNA *Reoviridae* like virus found in a *Odonata* (*Odonatoptera*) RNA pooled library (Shi et al., 2016). NcV6 RNA segment 2 is 2,617 nt long, presenting a single ORF (18-2,513 nt coordinates) encoding a putative capsid protein of 93.7 kDa and 831 aa long. NcV6 CP is slightly related (E-value = 4e-11, pairwise identity 22%) to the putative minor capsid protein of Hubei reo-like virus 11. *Reoviridae* are multipartite viruses composed of ca. 10 dsRNA genome segments, nevertheless HrlV10 is reported to be an only bi-segmented virus composed by the RP and a second RNA genome segment encoding a minor structural protein. Again, here it is postulated that HrlV10 might reflect segment loss during the evolutionary history of an ancestral *Reoviridae*. Likewise, NcV6 appears to be conformed also only by two RNA segments. In consequence, I suggest that as the HrlV10 antecedent, NcV6 might be a bi-segmented reo-like virus (dsRNA). Future studies could expand the current limited sequences reference set associated to these cluster of viruses, which might lead, in turn, to the identification of highly divergent potential virus segments that should not be ruled out this soon, based on absence of evidence. FasTree phylogenetic trees derived from MUSCLE alignments of NcV6 RP and related viruses hints that NcV6 and HrlV11 forms a divergent cluster within the *Reoviridae*, supporting a potential assignment to a new clade of Reo-like viruses.

New strain of reported invertebrate virus associated to *N. clavipes*

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In addition to the tentatively new virus species described above, I found sequences corresponding to recently reported arthropod viruses. Wuhan fly virus 6 (WFV6) has been identified in a *Diptera* (*Insecta*) RNA pooled sample (Shi et al., 2016) consisting of several classified and unidentified fly species, hence the actual host of WFV6 remains elusive. WFV6 has been tentatively assigned by the authors to a Partiti-Picobirna superclade. An inspection of the corresponding sequences suggests it could be more associated to the *Partitiviridae* family of dsRNA viruses. I found virus sequences corresponding to a new strain of WFV6, which shares a nucleotide identity of 97.5 % to the reference sequence at the RNA 1 segment. The *N. clavipes* strain of WFV6 is 8 nt longer (2 nt at 5´and 6 nt at 3´region). Interestingly, of the 33 single nucleotide polymorphism among the sequences, 31 corresponded to the 3rd base position of the expected codon sequence, and the other 2, although were located at the first base of the codon did not generate amino acid changes. Therefore, the WFV6 refseq and *N. clavipes* strain of WFV6 share a 100 % identity at the RdRP protein level, which could be associated to a certain level of constraint that prevents sequence divergence that might affect the RP functional domain (RdRP_1; Pfam id = PF00680; 152-372 aa coordinates) (Supp. Figure 9). Despite being WFV6 a putative member of a bi-partite family of viruses, Shi et al (2016) did only report a single RNA segment corresponding to a putative RdRP. Besides the RdRP encoding genome segment, *Partitiviridae* are composed typically of a second RNA segment, which encodes a coat protein (Nibert et al., 2014). By sequence homology searches, I found evidence of a putative RNA segment 2 that could be assigned to the *N. clavipes* strain of WFV6. This RNA is 1,410 nt long, presenting a single ORF (41-1,291 coordinates) encoding a 416 aa putative coat protein related to that of Wuhan insect virus 23 (Partiti-Picobirna), sharing a 22% sequence identity (E-value = 2e-11). Phylogenetic analyses based on the RdRP of WFV6 and the *N. clavipes* strain, group them within a clade of newly reported invertebrate *Partitiviridae* related viruses. This clade, among the *Partitiviridae* family, is linked (but strongly divergent) to the *Gammapartivirus* genus, which is exclusively constituted by mycoviruses.

Hubei virga-like virus 11 (HvlV11) has also been recently reported by Shi et al (2016) associated to a *Diptera* (*Insecta*) RNA pooled sample, lacking a confirmed specific host within the fly library. HvlV11has been tentatively assigned to a virga-like superclade of ssRNA (+) *Virgaviridae* like viruses. The authors suggested that HvlV11 genome is 6,206 nt long, harboring four ORFs encoding a 1,065 aa long replicase and 3 structural proteins. Nevertheless, based on a

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

detected strain of HvlV11 I found associated to *N. clavipes*, I suggest that the refseq assembled strain corresponded to a truncated, partial version of the corresponding virus genome. The tentative ATG the authors postulate as a transcription start site at position 120 nt, is an internal methionine, and corresponds to a new tentative position 4,367 nt within the genome. The *N. clavipes* strain of HvlV11 (HvlV11-Ncs) is 10,433 nt long, harboring four ORFs, but with some predicted variants in comparison to the refseq sequence (Supp. Figure 8). ORF1 of HvlV11-Ncs (119-7,657 nt coordinates) encodes a putative replicase protein (RP) of 287 kDa and 2,481 aa long (1,416 aa longer than the reported refseq). Several distinct functional domains were predicted covering the RP, between the 63-435 aa, 1,083-1,262 aa, 1,644-1,888 aa and 2,023-2,460 aa coordinates, corresponding to an alpha-virus like viral methyl-transferasa (Pfam id PF01660), a S-adenosyl-L-methionine-dependent methyltransferase (Superfamily db id SSF53335), a (+) RNA virus RNA helicase (Pfam id PF01443), and a Tymovirus-like RdRP (RdRP_2 Pfam id PF00978) respectively. The methyltransferase domains additionally found in this extended version of the HvlV11 replicase are consistent with its Virga-like assignment, supporting the hypothesis that the refseq sequence corresponds to a truncated partial sequence of the virus. HvlV11-Ncs presents three more ORFs at position 7,566-7,979 nt, 8,021-9,796 nt, and 9,843-10,322 encoding three potentially structural proteins of 137, 591 and 196 aa respectively. The uncharacterized protein encoded in ORF3 has no similarity to any other viral protein, but based on TMHMM searches, a putative transmembrane signal at the N-terminus was predicted (TMHMM2.0 id TMhelix, 29-51 aa coordinates) which may be linked with a potential role in movement of this 591 aa protein. Both ORF2 and ORF4 present a TMV-like viral coat protein domain (InterPro id IPR001337), and a TMV-like_coat (Pfam id PF00721) suggesting a role in coating of the viral genomic RNA, and supporting the assignment of HvlV11-Ncs as a virga-like virus.

Rehmannia mosaic virus (RMV) is a member of the plant infecting genus *Tobamovirus,* corresponding to the *Virgaviridae* family. RMV was first discovered from the traditional medicine associated herb *Rehmannia glutinosa* in China (Zhang et al., 2008). RMV infection elicits systemic mosaic symptoms in Rehmannia and is very similar to (perhaps an isolate of) *Tobacco mosaic virus* (TMV) the type *Tobamovirus* species. Only in one of the analyzed spider RNA libraries, a tentatively new strain of RMV associated to *N. clavipes* was found (Ncas-RMV). This cautious report of this finding should be interpreted only as a unconfirmed link between the RNA data and the potential of this virus to be a *bona fide* N. clavipes strain of RMV or a perhaps a false positive

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated to contamination, or other prospective sources. The Ncas-RMV was detected in only one of the silk glands derived samples. The corresponding sequence corresponds to a full length genome of RMV, conserving expected ORFs, sequence structure and domain architecture. Ncas-RMV shares a 93.2% and 97.3% sequence identity at the nt genome and at the 183 kDa replicase protein aa, respectively (Supp. Figure 10). Ncas-RMV replicase presented the typical TMV-like domains between the 50-466 aa, 830-1,085 aa, and 1,168-1,607 aa coordinates corresponding to an alpha-virus like viral methyl-transferasa (Pfam id PF01660), a (+) RNA virus RNA helicase (Pfam id PF01443), and a RdRP (RdRP_2 Pfam id PF00978) respectively. Ncs-RMV ORF2 and ORF3 encode a characteristic 30kDa movement protein and a 17.5 kDa coat protein presenting a TMV-like_coat domain (Pfam id PF00721). Nevertheless, based on the corresponding SRA Raw RNA data, the Ncas-RMV sequence is supported only by 914 reads (1.44 FPKM, mean coverage 14.1)(Library: 49.9 Million QC Filtered Read Pairs, Illumina HiSeq2000 (100 x 100)) (Supp. Figure 11). Ncas-RMV was the only virus sequence detected merely in a single independent RNA library, and with a strongly low mean coverage. Thus, its association to *N. clavipes* is weak. A borderline *ad hoc* hypothesis could be that during abdomen microdissection to reach this specific silk gland for sample preparation, there was a concomitant non target purification of insect prey harboring the virus, or plant tissue debris that allowed its sequencing in the corresponding library. It is important to highlight that TMV has no true vectors, although there have been reports of its transmission by aphids (Hemiptera), probably by mechanical means (Lojek et al., 1969). Moreover, TMV is very persistent on clothing and on glasshouse structures (Broadbent & Fletcher, 1963). In this scenario, the association between this RMV strain and *N. clavipes* should be confirmed in future studies. It is worth mention that every detected potential virus sequence was assessed on the whole genome assembly of *N. clavipes* (NepCla1.0; GenBank accession GCA_002102615.1; Babb et al., 2017) and no evidence that the tentative viruses could be derived from integration of virus-related sequences into the genome were found. In addition, given the fact that the detected sequences corresponded to the full length of the putative virus, present unaltered encoding and spacer regions and maintain the typical domain architecture of related viruses support that the identified sequences correspond to *bona fide* extant *N. clavipes* viruses. Moreover and importantly, in the case of multi-segmented nature viruses (*Partitiviridae*, *Reoviridae*), all the corresponding and expected segment RNAs were found.

Tissue/Organ presence and RNA levels of *N. clavipes* viruses

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The *N. clavipes* multiple RNA-seq data derived from 16 different tissue/organ/individual isolates (whole body, brain, and individual silk and venom glands) collected from four female individuals (Babb et al., 2017) allowed for the first time to address the presence of a spider virus RNA at the tissue/organ level (Figure 2; Supp. Figure 12; Supp. Table2). Virus presence and levels were assessed by fast read mapping using the Geneious assembler with low sensitivity and no iteration of the RNA-seq QC Filtered RNA data available at SRA. Results were normalized and expressed as FPKM, mean coverage calculated, variants/polymorphism estimated among samples, and the tentative virus sequences were curated based on base frequency. Virus transcript presence and levels were complex, consistent and varied by species, individual and tissue/organs assayed. A total 4,428,279 absolute reads were assigned to be derived from viruses. Essentially, virus derived RNA were retrieved on every sample, and to my knowledge this is the first time that a virus derived nucleic acid is detected specifically in silk and/or venom glands, and on the brain of spiders. If virus presence is estimated at the individual spider level, NcV1 and NcV2 were conclusively detected in every female spider sample, at different transcript levels, which varied in relation with the sampled organ origin. NcV3 and NcV6 were detected in most spiders but not in Nep-7. NcV4 and NcV5 were detected in both Nep-7 and Nep-9. HvlV11 (Ncs) was detected in Nep-8 and Nep-9, WFV6 (Ncs) only in Nep-9 and, and RMV (Ncs) only in a specific silk library of Nep-9. Notably, the detection and RNA levels of bipartite predicted viruses (NcV6 and WFV6) were consistent for the two corresponding RNA genome segments among samples, independently confirming virus presence on selected libraries. It is important to highlight that the RNA virus estimated loads differed significantly among samples. For instance, NcV1 levels were relatively high among most tissue samples but not on brain samples, on the contrary NcV6 levels were relatively low among samples, but strikingly spiked specifically on brain tissues, as is the case for NcV2 on the Nep-8 sample. In general, as a whole, the presence of virus RNA was significantly accumulated in additional magnitude at the brain tissue, ascending in one sample to a striking ca. 2.4% of total detected RNA reads (Figure 2.B-D). The biological significance of this finding, though interesting, remains elusive. Furthermore, the diverse viruses were consistently detected in the independent silk samples corresponding to the same spider, at a similar FPKM level, suggesting the accumulation of viruses is more-less steady among the diverse silk glands. Interestingly, when whole spiders were sampled, the detected viruses were found to be at lower RNA levels than in the specific silk/venom/brain libraries (Figure 2.A-B). Although in the context

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of a small sample size and the fact that the whole body libraries derive from different sampled individuals than the tissue libraries, I cautiously speculate that perhaps these specific sampled tissues are enriched on RNA virus loads. More complex distribution/presence absence patterns might be observed in Supp. Figure 12 and Supp. Table 2.

Conclusions

Regardless of sample size, and that the number of detected viruses is limited, it is interesting to highlight that most detected sequences corresponded to new unreported virus species. Spiders could be an important reservoir of viral genetic diversity that ought to be assessed. Widespread consistent RNA accumulation of diverse putative viruses, on independent profiled samples, sequence structure and domain architecture, supports the assumption that the identified sequences correspond to *bona fide* viruses. It is not easy to speculate about the biological significance of the presence, accumulation, and distribution of these potential viruses in the context of limited literature. The brain enrichment of RNA virus loads appears not to be incidental, and could be associated to a potential effect on the spider host. The accumulation of viral RNA on silk and venom glands may have some evolutionary relation with virus horizontal transfer. Future studies should focus not only on complementing and expanding these preceding findings, but also on addressing the potential ecological role of these viruses, which might influence the biology of these outstanding arthropod species.

Data availability.

All relevant data used as input has been generated and described in Babb et al (2017). Data from Babb et al (2017) are available through the central BioProject database at NCBI under project accession PRJNA356433 and BioSamples accessions SAMN06132062–SAMN06132080. All short-read sequencing data are deposited in the NCBI Short Read Archive (SRX2458083–SRX2458130), and transcriptome data are available at the Transcriptome Shotgun Assembly (TSA) under accession GFKT00000000.

Acknowledgements

I would like to express sincere thanks to Dr. Benjamin Voight for his encouragement in communicating these findings, which are only possible thanks to the remarkable work of his team and colleagues. Additional thanks to Dr. Casey Greene and colleagues for taking from words to

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

action the active support and promotion of secondary analysis of data, which could redound in new discoveries and hypotheses, and a paradigm shift in research practices.

References

-Agnarsson, I., Kuntner, M. & Blackledge, T.A. (2010). Bioprospecting finds the toughest biological material: extraordinary silk from a giant riverine orb spider. *PLoS ONE, 5,* e11234

-Babb, P. L., Lahens, N. F., Correa-Garhwal, S. M., Nicholson, D. N., Kim, E. J., Hogenesch, J. B., ... & Voight, B. F. (2017). The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. *Nature Genetics*, doi:10.1038/ng.3852

-Broadbent, L., & Fletcher, J. T. (1963). The epidemiology of tomato mosaic. *Annals of Applied Biology*, *52*:233-241.

-Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, *26* : 2620-2621.

-Olivier, V., Blanchard, P., Chaouch, S., Lallemand, P., Schurr, F., Celle, O., ... & Ribière, M. (2008). Molecular characterisation and phylogenetic analysis of Chronic bee paralysis virus, a honey bee virus. *Virus research*, 132: 59-68.

-Garrison, N. L., Rodriguez, J., Agnarsson, I., Coddington, J. A., Griswold, C. E., Hamilton, C. A., ... & Bond, J. E. (2016). Spider phylogenomics: untangling the Spider Tree of Life. *PeerJ*, 4, e1719.

-Ghosh, R. C., Ball, B. V., Willcocks, M. M., & Carter, M. J. (1999). The nucleotide sequence of sacbrood virus of the honey bee: an insect picorna-like virus. *Journal of General Virology*, 80:1541-1549.

-Govan, V. A., Leat, N., Allsopp, M., & Davison, S. (2000). Analysis of the complete genome sequence of acute bee paralysis virus shows that it belongs to the novel group of insect-infecting RNA viruses. *Virology*, 277: 457-463.

-Higgins, L. (2000). The interaction of season length and development time alters size at maturity. *Oecologia,* 122:51–59

-Higgins, L. E. & Buskirk, R. E. (1992). A trap-building predator exhibits different tactics for different aspects of foraging behaviour. *Animal Behaviour,* 44:485–499

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

-Kuntner, M. & Coddington, J. A. (2009). Discovery of the largest orbweaving spider species: the evolution of gigantism in *Nephila*. *PLoS ONE,* 4, e7516

-Li, C. X., Shi, M., Tian, J. H., Lin, X. D., Kang, Y. J., Chen, L. J., ... & Zhang, Y. Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife*, 4, e05378.

-Le Gall, O., Christian, P., Fauquet, C. M., King, A. M., Knowles, N. J., Nakashima, N., ... & Gorbalenya, A. E. (2008). Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T= 3 virion architecture. *Archives of virology*, 153: 715.

-Lojek, J. S., & Orlob, G. B. (1969). Aphid transmission of tobacco mosaic virus. *Science*, 164: 1407-1408.

-Nibert, M. L., Ghabrial, S. A., Maiss, E., Lesker, T., Vainio, E. J., Jiang, D., & Suzuki, N. (2014). Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research. *Virus research*, 188:128-141.

-Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., ... & Zhang, Y.Z. (2016). Redefining the invertebrate RNA virosphere. *Nature*, 540:539-543.

-Shean, R. C., Makhsous, N., Crawford, R. L., Jerome, K. R., & Greninger, A.L. (2017). Draft Genome Sequences of Six Novel Picorna-Like Viruses from Washington State Spiders. *Genome announcements*, *5*(9), e01705-16.

-Vollrath, F. (2000). Strength and structure of spiders' silks. *Reviews in Molecular Biotechnology*, 74:67-83.

-Zhang Z.C., Lei C.Y., Zhang L.F., Yang X.X., Chen R., Zhang D.S. (2008). The complete nucleotide sequence of a novel *tobamovirus*, Rehmannia mosaic virus. *Archives of Virology*, 153:595–599.

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

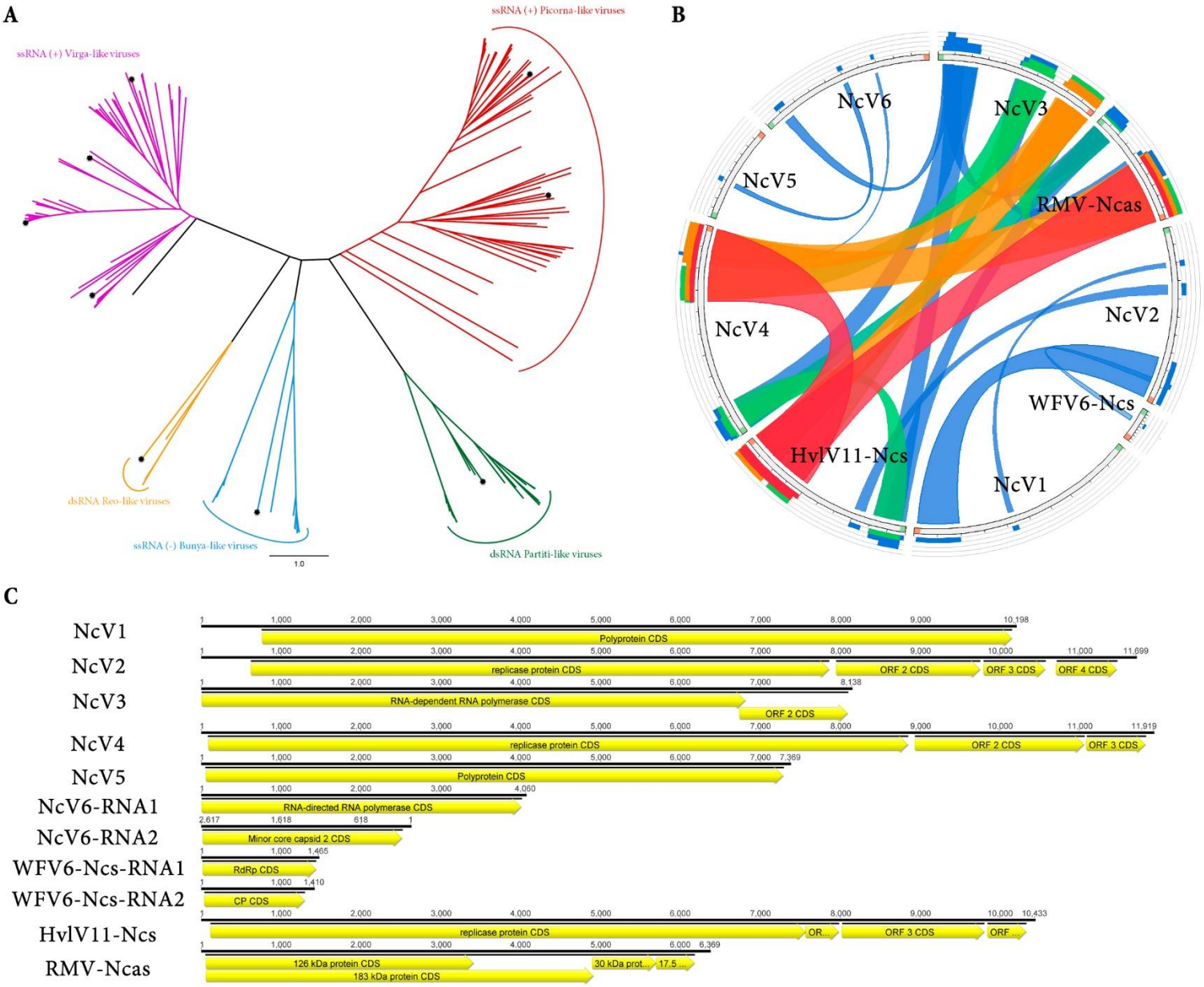

Figure 1. A) Maximum likelihood unrooted phylogenetic tree based in MAFFT alignments of predicted replicase proteins of *Nephila clavipes* associated viruses (black stars) and related reported viruses. Scale bar represents substitutions per site. Only specific families associated to the predicted viruses are represented for clarity. A complete tree showing tip species labels and FasTree support values is available as (Supp. Fig. 1). B) Sequence similarity levels of detected viruses expressed as Circoletto diagrams (Darzentas, 2010) with default parameters and based in BLASTP searches with an E-value of 1e-1 threshold. Replicase proteins (RP) are depicted clockwise, and sequence similarity is visualized from blue to red ribbons representing low-to-high sequence identity. The region with highest identity among RPs corresponds to the predicted RdRP associated functional domain located at the 3´region of the respective proteins. C) Genome graphs depicting RNA predicted genome or genome segment sequences of detected *N. clavipes* viruses. Abbreviations: NcV1-6, *Nephila clavipes* virus 1-6; WFV6-Ncs, *N. clavipes* strain of Wuhan fly virus 6; HvlV11-Ncs, *N. clavipes* strain of Hubei virga-like virus 11; RMV-Ncas, *N. clavipes* associated strain of *Remania mosaic virus*.

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

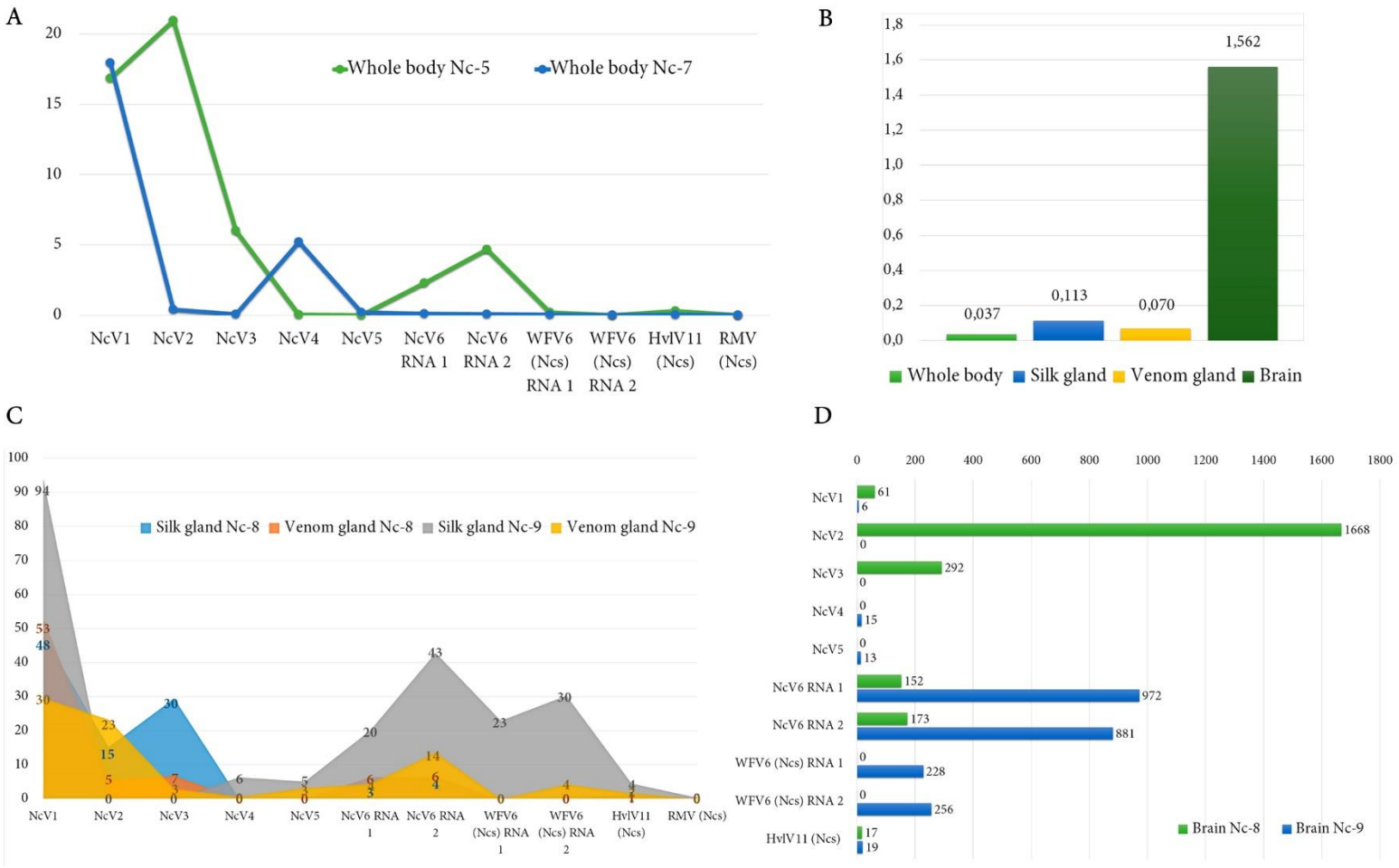

Figure 2. Graphs expressing RNA transcript levels assayed in two whole body spider samples, ten individual silk glands, two venom glands, and two brain isolates collected from four non-gravid females, Nep-5, Nep-7, Nep-8 and Nep-009. Values are expressed as FPKM in A, B, C, and as percentage of total QC filtered read pairs mapped to viruses in D. In C tissue FPKM are an average of biological replicates corresponding to each tissue. Values corresponding to each sample are available as Supp. Table 2.

bioRxiv preprint doi: https://doi.org/10.1101/140814; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. *Nephila clavipes* viruses, genome composition, similarity to reported viruses and proposed names.

Accesion #	Gen size	RP	Best hit	Bitscore	Query C	E-value	Identity	Proposed name
GFKT010682328.1	10198	3128	Wuhan spider virus 2	2175	97%	0.0	40%	*Nephila clavipes* virus 1
GFKT010485529.1	11695	2412	Wuhan spider virus 6	462	72%	4,00E-73	36%	*Nephila clavipes* virus 2
GFKT011117165.1	8138	2266	Lodeiro virus	836	97%	0.0	29%	*Nephila clavipes* virus 3
GFKT013490456.1	11919	2917	Shayang virga-like virus	1130	64%	0.0	34%	*Nephila clavipes* virus 4
GFKT013533247.1	7369	2409	Andes orthohantavirus	112	35%	8,00E-21	22%	*Nephila clavipes* virus 5
GFKT013780701.1	4060	1331	H. vitripennis reovirus	129	70%	3,00E-26	23%	*Nephila clavipes* virus 6 RNA 1
GFKT013802461.1	2617	831[a]	Hubei reo-like virus 11	104	50%	2,00E-05	24%	*Nephila clavipes* virus 6 RNA 2
GFKT013514195.1	1465	470	Wuhan fly virus 6	983	100%	0.0	100%	Wuhan fly virus 6 (N. clavipes strain) RNA 1
GFKT013832808.1	1410	416[b]	Wuhan insect virus 23	69.7	65%	4,00E-11	22%	Wuhan fly virus 6 (N. clavipes strain) RNA 2[c]
GFKT013930511.1	10433	2481[d]	Hubei virga-like virus 11	2189	42%	0.0	99%	Hubei virga-like virus 11 (N. clavipes strain)
GFKT013532199.1	6367	1615	*Rehmannia mosaic virus*	1161	96%	0.0	97%	*Rehmannia mosaic virus* (N. clavipes ass. strain)

Abbreviations: Gen size, RNA predicted genome size in nucleotides; RP gene product size in aa corresponding to predicted replicase proteins. Best hit, closest hit on NCBI BLASTP searches with predicted gene product against the non-redundant protein database NR; Query C, query coverage. [a] corresponds to a predicted structural protein of NcV6. [b] corresponds to the predicted (unreported) coat protein gene product of WFV6-Ncs. [c] This genome segment has not been described in the literature. [d] The refseq accession NC_033082.1 is a truncated 6,206 nt long partial sequence of HvlV11.