nature communications

Article

https://doi.org/10.1038/s41467-025-61035-2

Discovery of novel non-retroviral endogenous viral elements reveals their long-term integration history in spiders

Received: 29 August 2024

Accepted: 6 June 2025

Published online: 01 July 2025

Check for updates

Hengrui Hu[1], Just M. Vlak[2], Zhihong Hu[1] & Manli Wang[1]

Endogenous viral elements (EVEs) are widespread in the genomes of various organisms and have played a crucial role in evolution. Historically, research on EVEs primarily focused on those derived from retroviruses; however, the significance of non-retroviral EVEs (nrEVEs) has gradually gained recognition. In this study, we employ an approach that combines protein structure prediction with sequence analysis to identify a large group of previously unrecognized nrEVEs across spider genomes. Additionally, we identify nrEVE-related messenger RNAs, small interfering RNAs, and PIWI-interacting RNAs in spiders, suggesting that these nrEVEs may be functionally active. We also experimentally confirm the presence of spider nrEVEs and their transcripts in individual spiders. Evolutionary analysis suggests that these spider nrEVEs derived from unidentified nuclear arthropod large DNA viruses belonging to the order *Lefavirales*, class *Naldaviricetes*. Multiple integration events must have occurred both anciently and recently during the evolutionary history of spiders to explain these nrEVEs. Our findings reveal a novel group of nrEVEs and provide valuable insights into their evolutionary relationship with arthropods.

The genomic integration of viral elements has been reported across all three domains of life[1,2]. The genomes of organisms commonly harbor transposons and endogenous viral elements (EVEs), and retroviruses have historically been recognized as the major promoters of EVE integration[3]. Recent advances in genome sequencing and bioinformatics have revealed numerous non-retroviral endogenous viral elements (nrEVEs) in various organisms[4,5]. The mechanisms of nrEVEs integration remain largely unclear; however, nrEVEs evidently undergo selection by the host, who sometimes exploit them for various functions[6,7]. These apparent remnants of ancient viral infections, which serve as molecular fossils, provide valuable insights into the evolution dynamics between hosts and viruses over evolutionary timescales[2,8,9].

Arthropods comprise the most diverse group of animals on Earth and include two major subphyla: Mandibulata (e.g., insects and crustaceans) and Chelicerata (e.g., horseshoe crabs, scorpions, and spiders)[10]. Arthropods host a diverse and complex array of viruses owing to their rich species diversity. Among these, a distinctive group of viruses, known as nuclear arthropod large DNA viruses (NALDVs), are exclusively distributed in insects and crustaceans. NALDVs, now classified within the class *Naldaviricetes*, possess circular double-stranded DNA genomes ranging from 80 to 500 kb, and replicate within the host cell nucleus. They encompass four identified families: *Baculoviridae*, *Hytrosaviridae*, *Nimaviridae*, and *Nudiviridae*[11], and recently proposed *Filamentoviridae*[12], as well as the unclassified Apis mellifera filamentous virus (AmFV). Hundreds of NALDV species have been isolated, forming a monophyletic group separate from other arthropod large DNA viruses classified as nucleocytoplasmic large DNA viruses (NCLDVs) within the phylum *Nucleocytoviricota*[12]. The virion envelope protein known as peroral infectivity factors (PIFs), unique core proteins identified solely in *Naldaviricetes* and bracoviruses of *Polydnaviriformidae*, are shared across all NALDVs[11]. Within the class *Naldaviricetes*, *Baculoviridae*, *Hytrosaviridae*, *Nudiviridae*, and *Filamentoviridae* are further grouped to collectively form the order *Lefavirales* as they all share the so-called late expression factors (LEFs)[13].

[1]State Key Laboratory of Virology and Biosafety, Wuhan Institute of Virology, Chinese Academy of Science, Wuhan, China. [2]Laboratory of Virology, Wageningen University, Wageningen, The Netherlands. e-mail: huzh@wh.iov.cn; wangml@wh.iov.cn

nrEVEs originating from NALDVs have been identified in several arthropods and represent a rare integration of large DNA viral genome segments into eukaryotic genomes. Genomic integration of nudiviruses into cultured cells has been reported and fossilized remnants of nudiviruses have been detected in the genomes of numerous arthropods[14,15]. Bracoviruses are derived from an endogenous nudivirus within the wasp genome[16] and exhibit remarkable symbiotic relationships with wasp hosts. Bracoviruses are essential for wasps, as they can suppress the host immune response and create a suitable environment for wasp larval development[17]. Endogenous nimaviral genomes were discovered in host crustacean genome[18], and endogenization and domestication of filamentoviruses and AmFV were identified in multiple hymenopterans[19]. In summary, the nrEVEs of NALDVs appear to be common feature.

Spiders, classified as chelicerates (Subphylum Chelicerata), diverged from common ancestors of insects and crustaceans early in evolutionary history[20]. Currently, more than 50,000 documented spiders (Order Araneae) are members of two suborders, Mesothelae and Opisthothelae; the latter contains two infraorders, Mygalomorphae and Araneomorphae[21]. Members of the suborder Mesothelae form a sister group to all other living spiders, as they retained ancestral characters, such as a segmented abdomen with spinnerets in the middle and two pairs of book lungs. Spiders in the infraorders Mygalomorphae, including tarantulas and trapdoor spiders, are robust and "hairy," with large chelicerae and fangs. The Araneomorphae lineage, also called "true spiders," comprises the vast majority (~93%) of living spiders, including orb-web spiders, wolf spiders, and jumping spiders[21]. Research on spiders has predominantly focused on silk and venom[22]. Only a limited number of RNA viruses and EVEs have been identified in spiders[23], and their impact on spider populations and ecosystems remains unknown. No NALDV nor NCLDVs have been identified in spiders to date.

Conventional methods of EVEs discovery rely primarily on homologous sequence searches for known viruses within the host species. Hence, EVEs with low similarity to known viral sequences are unlikely to have been discovered. Recent, advancements in artificial intelligence technology, particularly in protein structure prediction and comparison, have offered new avenues for the discovery and classification of distantly related viral sequences and previously unknown protein components[24–26]. In this study, we combined protein structure prediction and sequence alignment to identify baculovirus-related sequences in databases and uncover novel nrEVEs in spider genomes. We further investigated the distribution of these nrEVEs in spider genomes and uncovered the messenger RNAs (mRNAs), PIWI-interacting RNAs (piRNAs), and small interfering RNAs (siRNAs) related to these nrEVEs in the available transcriptional data on spiders. We also experimentally verified the presence of nrEVEs and their transcripts in the spider samples. Finally, we reconstructed the evolutionary relationships between spider nrEVEs, NALDVs, and their host arthropods.

Results

Discovery of abundant baculoviral homologs in spider genomes using a combined strategy of structure and sequence comparisons

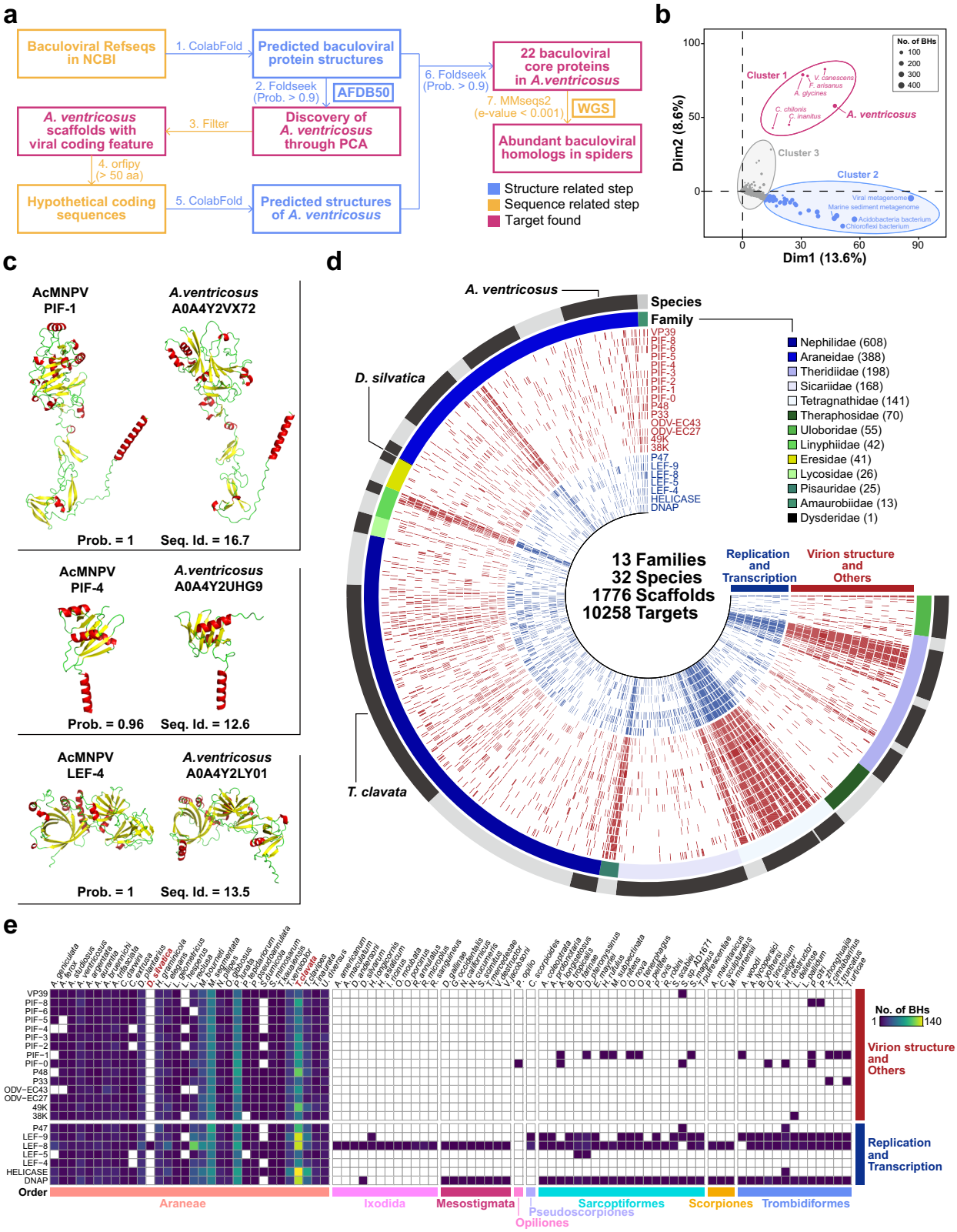
We developed a comprehensive approach (Fig. 1a) to identify the uncovered NALDVs nrEVEs. First, by querying the predicted baculoviral protein structures in the AlphaFold Protein Structure Database cluster (AFDB50), we identified numerous species harboring multiple baculoviral homologs (probability over 0.9) (Supplementary Data 1). Via principal component analysis (PCA) of the species, we identified a distinct spider species, *Araneus ventricosus*, clustered with four parasitoid wasps (*Cotesia chilonis*, *Chelonus inanitus*, *Fopius arisanus*, and *Venturia canescens*) and an aphid (*Aphis glycines*), all of which were reported to contain NALDV-like integrations[16,27–29] (Fig. 1b). We therefore focused on analyzing the spider genome in more detail. Structural alignment revealed significant similarities between the proteins identified in *A. ventricosus* and their baculoviral homologs. For example, three proteins showed a homology probability over 0.95, with the conserved baculovirus core proteins (PIF-1, PIF-4, and LEF-4) of Autographa californica multiple nucleopolyhedrovirus (AcMNPV), while their amino acid sequence identities were below 20% (Fig. 1c). Subsequently, the open reading frames (ORFs) of selected scaffolds (Supplementary Data 2) with large DNA virus-coding strategies (e.g., higher ORF density) were extracted, and their structures were predicted. After structural comparison and sequence alignment of these proteins with the predicted baculoviral protein structures, we identified over 30 distinct baculoviral homologs, including 22 baculoviral core proteins (conserved among all baculoviruses) (Fig. 1a). These core proteins included those involved in DNA replication and transcription (e.g., Helicase, DANP, P47, LEF-4, LEF-5, LEF-8, and LEF-9), as well as virion structure and other functions (e.g., PIF-0, PIF-1, PIF-2, PIF-3, PIF-4, PIF-5, PIF-6, PIF-8, VP39, 38K, 49K, ODV-EC43, ODV-EC27, P33, and P48) (Supplementary Data 2).

To comprehensively investigate the occurrence and distribution of baculoviral homologs in spider genomes, we screened 22 baculoviral core protein homologs of *A. ventricosus* against all available arachnid genome assemblies in the National Center for Biotechnology Information (NCBI) Whole Genome Shotgun (WGS) database, encompassing 92 species from diverse orders, including Araneae, Trombidiformes, Sarcoptiformes, Ixodida, Mesostigmata, Scorpiones, Opiliones, and Pseudoscorpiones (Supplementary Data 3). Our analysis revealed 10641 baculoviral homologs across 89 species (Supplementary Data 4). Among the 43 assembled genomes from 32 different spider species, 10258 baculoviral homologs were identified in 1776 scaffolds, spanning 13 distinct families (Fig. 1d), with 12 families from the infraorder Araneomorphae and one family (Theraphosidae) from more ancient infraorder Mygalomorphae. Most baculoviral homologs were identified in the order Araneae, whereas a few baculoviral homologs with lower identity levels were identified in other orders (Fig. 1e). Interestingly, baculoviral homologs were identified in all the 32 genome-sequenced spider species, although only one homolog (LEF-8) was found in *Dysdera sylvatica* (Fig. 1e).

Broad distribution of baculovirus-related nrEVEs in spider genomes

Spider scaffolds harboring baculoviral homologs (Supplementary Data 5) displayed a wide spectrum, ranging from several kilobase pairs to hundreds of millions of base pairs (Fig. 2a and Supplementary Fig. 1a). Among the 1776 identified scaffolds in the WGS database, 927 scaffolds (52.20%) contain only a single baculoviral homolog, suggesting these nrEVEs are widely distributed in the spider genomes (Supplementary Fig. 1b). Interestingly, a small portion of scaffolds (6.19%) contain homologs of all 22 baculoviral core proteins (Supplementary Fig. 1b). Several species containing hundreds of baculoviral homologs were further analyzed (Fig. 2b). The entire chromosomes of *T. clavata* from SpiderDB (Supplementary Data 6), and selected scaffolds of *Hylyphantes graminicola*, *Latrodectus elegans*, *Dolomedes plantarius*, *Meta bourneti*, *Metellina segmentata*, and *Parasteatoda lunata* from the WGS database contain clustered distribution of baculoviral homologs. Chromosomal regions with baculoviral homologs usually contained direct repeats, inverted repeats, large fragments of repetitive sequences, and host genes, confirming that these baculoviral homologs were endogenous in the spider genome (Fig. 2c). Compared to the adjacent regions, the baculoviral homology rich regions displayed a notable increase in the percentage of ORFs coverage and GC content, suggesting sequential integration of large viral DNA fragments[30,31] (Fig. 2c). Collectively, the distribution and localization of baculoviral homologs confirmed the widespread existence of baculovirus-related nrEVEs in spider genomes.

https://doi.org/10.1038/s41467-025-61035-2



Evidences for pseudogenization

Pseudogenization is a common feature of endogenization. We analyzed the integrity of baculovirus-related genes in all the genome-sequenced spiders by query coverage (fraction of the query sequence covered by the target alignment). As shown in Fig. 3a, query coverage of most of the baculovirus-related genes in spiders was low, indicating they are pseudogenes. The number of pseudogenes is particularly high in *T. clavata*, which may contribute to the highest number of baculoviral homologs identified in this species (Fig. 2). The pseudogenes usually have multiple premature termination sites and appear more frequently in longer genes such as *dnap*, *helicase*, and *lef-8* as shown in the representative scaffolds of the indicated species (Fig. 3b). An extreme example is present in the scaffold BMAO01021503.1 of *T. clavata*, where the helicase gene mutated and evolved over time into nine pseudogenes.

Fig. 1 | Abundant baculoviral homologs in Araneae genomic data. a Survey strategy and flowchart illustrating the process of discovering baculoviral homologs (BHs) in arachnids. Different search methods and targets founded are represented by different colors and marked; the numbers indicate the survey order. The thresholds of different survey strategy are labelled alongside. **b** PCA result of different taxon within AFDB50 harboring baculoviral homologs. The result was binned into three clusters using K-mean analysis. The Cluster 1 includes a spider species (*Araneus ventricosus*), four parasitoid wasps (*Cotesia chilonis*, *Chelonus inanitus*, *Fopius arisanus*, and *Venturia canescens*) and an aphid (*Aphis glycines*), all of which reported to contain NALDV integrations except *A. ventricosus*. **c** Structures identified in the AFDB50 from *A. ventricosus*, along with their homologous proteins from AcMNPV. The structures are color-coded based on secondary structure. The estimated probability for query and target to be homologous (Prob.) and their sequence identity (Seq. Id.) from Foldseek are labeled below. **d** Diagram shows the scaffolds containing BHs in the WGS database from the order Araneae. The outer rings represent information on the families and species. The target number of scaffolds from each family are presented on the side. All the scaffolds from single species were concatenated. The diagram was visualized via anvi'o. **e** BHs identified in genomes of different species in the class Arachnida are shown in a heatmap. The classification of species is labeled below the heatmap. The results in (**d** and **e**) obtained by querying 22 distinct homologs of baculovirus core proteins identified in *A. ventricosus* using MMseqs2 against WGS database. The homologs of 22 distinct baculoviral core proteins are classed into two group based on their function.

Spider nrEVEs-related mRNAs, piRNAs and siRNAs exist in spiders

We further conducted an extensive survey of the baculoviral homologs in the NCBI Transcriptome Shotgun Assembly (TSA) database. In total, 1689 TSA datasets covered 1138 different spider species from 76 families (Supplementary Data 7). Remarkably, mRNAs with baculoviral homologs were widespread in Araneae (Supplementary Data 8). In addition, the number of baculoviral homologs in different spider families was strongly correlated with the number of datasets available for each family (Fig. 4a). However, the detection rate of core baculoviral homologs varied among the different families, and no obvious correlation was identified between the detection ratio and spider classification (Supplementary Fig. 2a).

We further examined the transcriptomic data of *T. clavata* available in SpiderDB, which contains the most detailed data for different spider tissues. Nearly full-length mRNA of certain baculoviral homologs were recovered (Supplementary Fig. 2b). Although the expression levels of EVE-related mRNAs were relatively limited (FPKM < 3), they displayed a broad expression pattern in the transcriptome data from different spider tissues, with a higher detection ratio in different parts of the major silk gland (Supplementary Fig. 2c).

Additionally, we surveyed the small RNA (sRNA) sequencing data of the major silk glands of *T. clavata*, revealing highly abundant sRNAs with many baculoviral homologs. Notably, the piRNA levels of *odv-ec43*, *p48*, and *p47* were significantly higher in the major silk glands, and the sRNA patterns of *pif-4*, *lef-9*, and *38k* showed a typical siRNA pattern (Fig. 4b). In addition, piRNAs appeared to be mainly present in the ducts of major silk glands, whereas siRNAs were mainly present in the tail (Fig. 4b).

Identification of spider nrEVEs and related mRNAs in different spider individuals

To further confirm the prevalence of nrEVEs in individual spiders, we purified genomic DNA from the collected *A. ventricosus* spiders and subjected them to high-throughput sequencing. Sequencing quality was confirmed by mapping the reads to *A. ventricosus* reference assembly scaffolds (accession number: BGPR01), with an overall alignment rate of 82%. The alignment details of the top nine scaffolds (AvEVE1-9), which contained the highest numbers of the distinct baculoviral homologs, were presented in Table 1 and Supplementary Fig. 3. Notably, the reference genomic assembly of *A. ventricosus* was generated from an individual spider captured in Yamagata, Japan, while our sequencing data was generated with an individual spider captured in Nanyang, China. The variation of gene coverage in some AvEVEs, suggests that the pattern of the nrEVEs is not fixed in the population of spiders. These nrEVEs showed different gene organizations, suggesting they may have been derived from multiple integrations and recombination events (Supplementary Fig. 3). We also noticed substantial single nucleotide polymorphisms in these AvEVEs, and the ratio (Ka/Ks) of the non-synonymous substitution rate (Ka) and synonymous substitution rate (Ks) of intact ORFs between our AvEVEs sequences and the reference assemblies were analyzed. The result revealed that the ORFs in the AvEVEs underwent strong purifying selection[32] (Fig. 4c). Moreover, we performed PCR assays to identify the presence of baculoviral homologs (*lef-8*, *p33*, *pif-0*, and *pif-1* of AvEVE2 and *lef-9*, *pif-5*, *vp39*, and *38k* of AvEVE3) in the five *A. ventricosus* individuals. Remarkably, these genes were consistently detected in all the tested spiders, regardless of sex, indicating the existence of AvEVEs in *A. ventricosus* genome (Supplementary Fig. 4).

In addition, samples of *T. clavata* were used to evaluate the expression levels of nrEVEs genes in diverse tissues via quantitative reverse transcription PCR (qRT-PCR). The results demonstrated a markedly higher expression level of baculoviral homologs in the silk gland than in other tissues (Fig. 4d).

Spider nrEVEs likely derived from unidentified NALDVs

To understand the origin of spider nrEVEs, sequences of the conserved homologs in *Naldaviricetes*[11], PIFs and LEFs, were subjected to maximum likelihood phylogenetic analysis. The resulting phylogenetic trees of PIF-0 showed that spider nrEVEs formed a monophyletic group comprising three distinct classes (Classes I, II, and III), that were evolutionarily distant from other NALDVs (Fig. 5a). Assuming that nrEVE genes with relatively close vicinity in the same scaffolds are likely originated from the same virus, we conducted phylogenetic analyses using concatenated sequences of PIFs or LEFs from the same scaffolds. The phylogenetic tree based on concatenated sequences of PIFs and LEFs also showed that spider nrEVEs formed a monophyletic group with high confidence and that members of this group were evolutionarily distant from other NALDVs (Fig. 5b, c). Notably, phylogenetic trees based on concatenated sequences of PIFs or LEFs confirmed the separation of Class I and Class II of spider nrEVEs but lacked Class III (Fig. 5b, c). This is because in contrast with Class I and Class II nrEVEs, which contain closely distributed PIFs and LEFs, Class III typically contains isolated nrEVE sequences.

Among the 1776 scaffolds identified, 336 were of length between 80 and 500 kb. Of these, 62 scaffolds contained a complete set of LEFs or PIFs (Supplementary Fig. 5a). Further phylogenetic analysis revealed that 52 scaffolds carried Class I homologs, 9 scaffolds carried Class II homologs, and one scaffold (CAJRAJ020013330.1) carried both Class I and II homologs, representing 14 spider species across nine distinct families. To understand the origin of spider nrEVEs, we selected nine representative scaffolds (Supplementary Data 9). Among these, three scaffolds carried Class II homologs and six scaffolds carried Class I homologs; and they belong to seven spider species within five distinct spider families (Supplementary Fig. 5b). Limited number NALDV-related pseudogenes were observed in these scaffolds. All proteins from these nine scaffolds and representative NALDVs were analyzed via structural comparison (Supplementary Data 10). Based on our structural comparison results and previously studies[11,33,34], we systematically summarized the conserved proteins of spider nrEVEs and NALDVs (Supplementary Data 11 and Supplementary Fig. 6). Notably, these scaffolds collectively harbored all conserved NALDV and lefaviral proteins, except Ac81. In addition, they contained the most conserved baculoviral and nudiviral proteins, except for VLF-1 and P6.9, as well as

https://doi.org/10.1038/s41467-025-61035-2

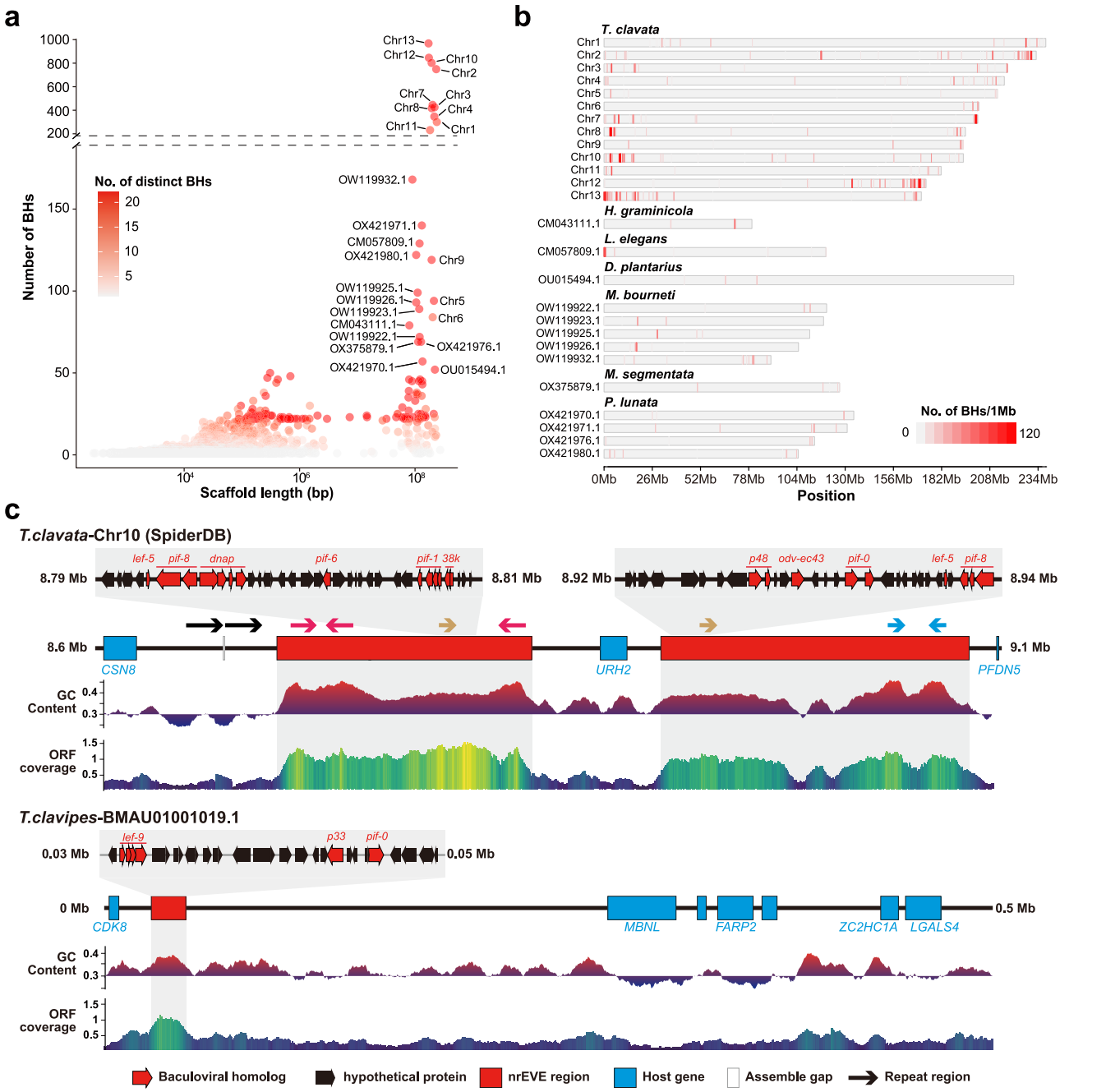


Fig. 2 | Distribution of baculoviral homologs in spider genomes. a This schematic illustrates the relationship between scaffold length and the number of baculoviral homologs (BHs) in WGS databases of Araneae and SpiderDB (Chr1-Chr13 of *T. clavata*). The color of the points represented the number of homologs of 22 distinct baculoviral core proteins in scaffolds. **b** The distribution of BHs in selected spider chromosomes or scaffolds. The lines represent the location of BHs, with the color shades indicating the number of BHs in 1 million base pairs (Mb) window size. The plot is generated using the center positions of BHs via R package 'CMplot'. **c** Two representative scaffolds containing host genes and BHs. The detail of nrEVE regions were enlarged above the scaffolds. The GC content and ORF coverage are showed below. The coverage of ORFs was calculated as the total ORFs length compared to the sequence length in 10 kb window sizes. The labels of feature types are displayed below the diagram. Identical repeat regions are shown with arrows of the same color.

some baculovirus-conserved proteins (49K, ODV-EC43, ODV-EC27, and P48) (Fig. 5d and Supplementary Data 11). Five proteins conserved in spider nrEVEs (helicase-2, flap endonuclease, DNA-binding protein, nudix hydrolases, and Ac106/107) were found in other NALDVs. Additionally, 11 predicted proteins were conserved in spider nrEVEs that did not exist in NALDVs (Fig. 5e), including homologs of protein kinases and innexins (Fig. 5f). Characterization of the coding regions of innexins suggested that they were derived from an integrated viral genome rather than from the host itself (Supplementary Fig. 7). Collectively, our data suggests that spider nrEVEs evolved from unclassified NALDVs belonging to the order *Lefavirales*, of the class *Naldaviricetes*.

Long-term integration history of nrEVEs in spiders

To understand the relationship between spider nrEVEs and their host spiders, we systematically conducted a co-phylogenetic analysis of the PIF-0 proteins with the classification of spiders. The results showed that the phylogenetic classification of spider nrEVEs correlated with the taxonomy of spiders (Fig. 6a and Supplementary Fig. 8). Specifically, the nrEVEs and transcripts from Classes I and II exclusively exist

https://doi.org/10.1038/s41467-025-61035-2

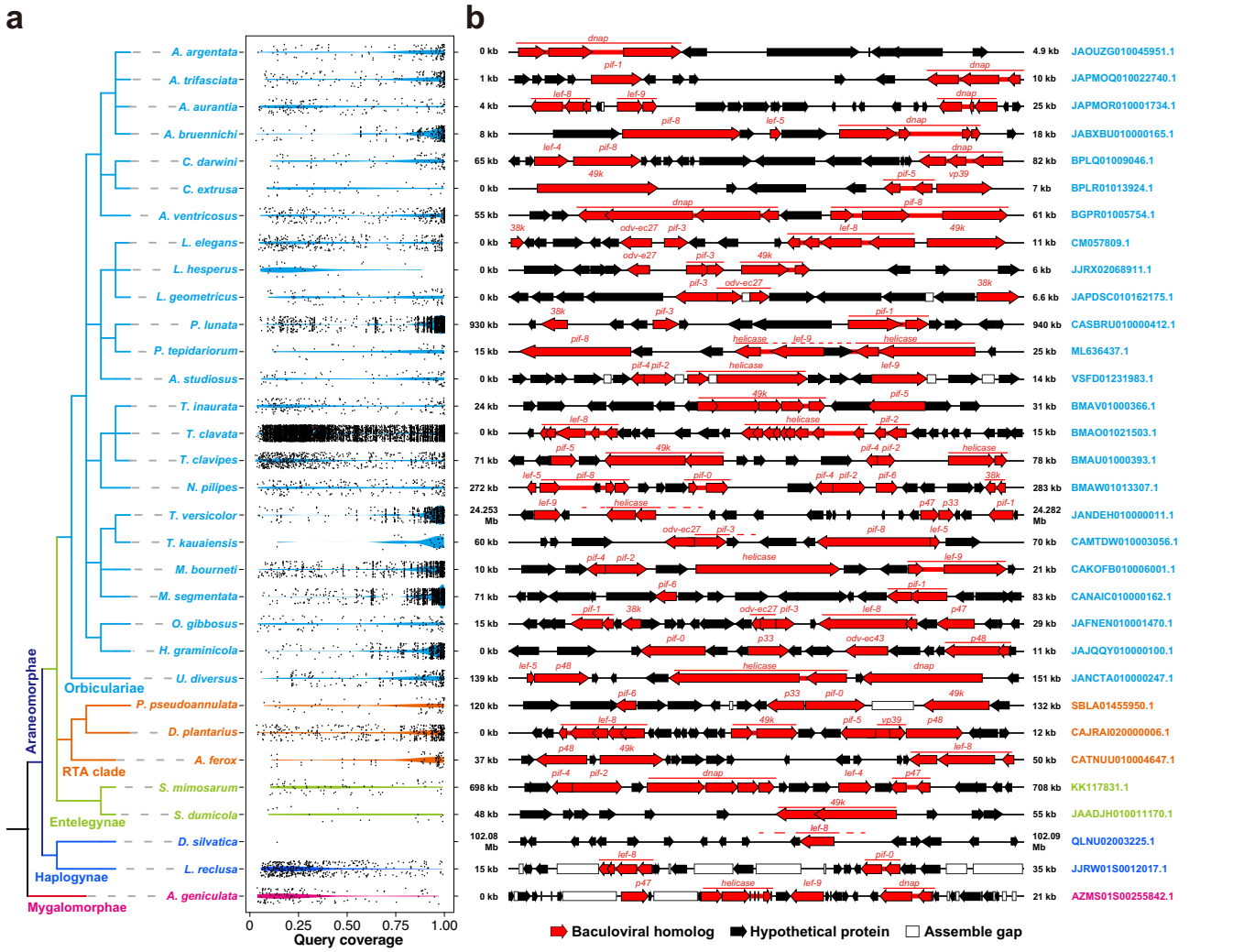


Fig. 3 | Detail characterization of baculoviral homologs in different spider species. a Different spiders are classified according to common tree from NCBI taxonomy and the classification was showed alongside. The dot plot shows the query coverage (fraction of query sequence covered by alignment) of baculoviral homologs in different spider species. **b** The diagram shows indicated scaffold containing baculoviral (pseudo)genes in the correspondent spider species with pseudogenes marked underlined. The labels of feature types are displayed below the diagram.

in the Entelegynae lineage, including the RTA clade, Orbiculariae clade, and other lineages (e.g., Palpimanoidea and Eresoidea). The nrEVEs from Class III exist in more ancient spider lineages, such as Mesothelae, Mygalomorphae, and Haplogynae, and in a few species from the RTA and Orbiculariae clades.

We performed molecular dating analyses to elucidate the evolutionary events and time scales of nrEVEs in spiders, and their evolutionary history with NALDVs. By using the time to the most recent common ancestor (TMRCA) of Chelonus inanitus bracovirus (CiBV) and Cotesia congregata bracovirus (CcBV)[35,36], and the fossil data of spiders[37–40] as a calibration point, we estimated the age of spider nrEVEs from three classes and other NALDVs using the concatenated sequences PIF-0 and LEF-8 (Fig. 6b and Table 2). Our analysis indicated that the emergence of the TMRCA of NALDVs was ~560 million years ago (Mya); the TMRCA of spider nrEVEs and nimaviruses were estimated to be ~497 Mya; the TMRCA of all lefaviruses were ~495 Mya; the divergence times of nudivirus and baculovirus was ~465 Mya, which is consistent with previous studies[35]; and the TMRCA of hytrosavirus, filamentovirus, and AmFV were ~458 Mya. Moreover, our molecular dating analysis revealed that the TMRCA of all spider nrEVEs were ~373 Mya, and the divergence time of the spider nrEVEs was associated with those of the species to which they belonged. In addition, the substitution rate of spider nrEVEs is 0.0036 substitutions per site per million years (subs/Mya), which is significantly slower than that of NALDVs themselves (0.0041 subs/Mya) ($P < 0.05$), supporting the view of endogenization of NALDVs (Fig. 6b).

Discussion

Research on EVEs has predominantly focused on retroviruses, and recently identified nrEVEs also primarily originate from RNA viruses[41,42]. Studies on nrEVEs from large DNA viruses have mainly concentrated on *bracoviriform*, which originate from nudivirus. Other NALDVs, such as hytrosavirus, filamentovirus, and nimavirus, have also been found to integrate into insect and crustacean hosts[11,43]. In the present study, we identified a significant number of novel nrEVEs in spiders using a combination of structural and sequence analyses (Fig. 1a). The discovery of spider nrEVEs has not only significantly broadened the known diversity of nrEVEs, but also expanded the host range of NALDVs to include two subphyla of arthropods, Chelicerata, and Mandibulata, indicating that ancient NALDVs may have widely infected all arthropods.

Based on the analysis of the distribution and evolutionary relationship of nrEVEs between spiders and NALDVs, it was speculated that the integration events of unidentified NALDVs might have happened

https://doi.org/10.1038/s41467-025-61035-2

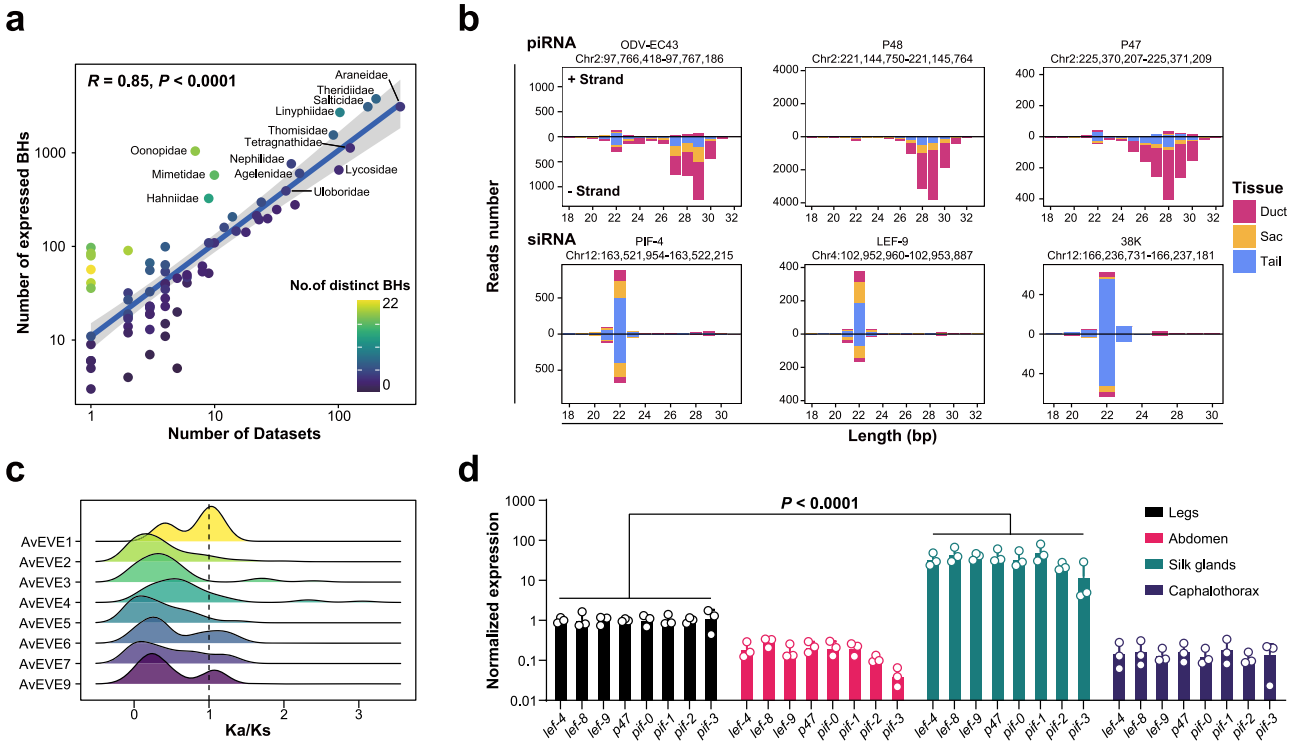


Fig. 4 | Identification and characterization of spider nrEVEs and related transcripts in different individual. a This scheme illustrates the relationship between the number of datasets and the identified number of expressed baculoviral homologs (BHs) in different Araneae families from TSA databases. The color of the points represented the number of homologs of 22 distinct baculoviral core proteins in scaffolds. **b** The size and strand distribution of small RNAs targeting BHs in three parts of major silk gland of available *T. calvata* sRNA sequencing data is depicted. The localization of BHs on chromosomes is labeled as a reference. **c** The Ka/Ks analysis of top nine scaffolds containing the highest number of distinct homologs of baculoviral core proteins (AvEVE1-9). The Ka/Ks analysis of AvEVE1-9 was carried out on the intact ORFs between our sequencing data generated consensus sequences and the reference genome assembly of WGS database (BGPR01). **d** The qRT-PCR result of BHs of collected female *T. clavata* samples ($n = 3$) in different tissues. The expression of BHs was first normalized with 16S rRNA and then compared to the corresponding expression levels of leg. Data represent mean ± s.e.m. Statistical analysis was performed using repeated-measures two-way ANOVA with multiple comparisons and Tukey's correction. The source data and *P* values were provided in Source Data file.

Table 1 | Alignment detail of AvEVE1-9

nrEVEs	Accession	No. of DBHs[a]	Length (bp)	No. of Reads[b]	Mismatch[c] (%)	Coverage[d] (%)
AvEVE1	BGPR01003736.1	20	128845	3926	3.1	42.62
AvEVE2	BGPR01006520.1	20	79615	18629	1	98.24
AvEVE3	BGPR01005754.1	18	89329	226558	1.1	87.76
AvEVE4	BGPR01001437.1	7	257312	28483	3.2	56.45
AvEVE5	BGPR01014530.1	7	39014	11448	1.3	96.90
AvEVE6	BGPR01017730.1	7	32001	18304	1.5	77.52
AvEVE7	BGPR01014537.1	6	39004	68159	1.6	95.13
AvEVE8	BGPR01022151.1	6	25695	1327	2.6	72.74
AvEVE9	BGPR01031596.1	5	17954	37204	1.7	70.70

[a]*DBHs* distinct baculoviral homologs.
[b]The read number was gained from the alignment result of sequencing data mapping to the full assemble scaffolds of BGPR01.
[c]The mismatch percent represents the mismatch rate of all reads.
[d]The coverage percent represents the percent of a scaffold is covered by reads.

multiple times in spiders, both anciently and recently (Fig. 7). The nrEVEs were found in all different suborders of spiders, while the most related order to spiders in our study, Scorpiones, was found to have an absence of related elements (Fig. 1e). The divergence time of Scorpiones and Araneae was ~397 Mya[44], and the molecular dating time of the TMRCA of spider nrEVEs was ~373 Mya (Fig. 6b). Therefore, the initial integration event might have occurred during the Devonian periods and after the diversion of Scorpiones and Araneae.

The baculoviral homologs in the more ancient suborder Magalomorphae and the clade Haplogynae are exclusively of the Class III type (Fig. 6a), and their integrity is relatively lower (Fig. 3a). This suggests that they are likely derived from an ancient integration event(s) and have undergone through extensive endogenization. Different to Class III, the Classes I and II nrEVEs are exclusively found in the RTA clade and the clade Orbiculariae (Fig. 6a). The molecular dating analyses confirmed that Class III was evolved earlier than the Classes I and II nrEVEs (Fig. 6b). In addition, a small portion of identified scaffolds contained nearly complete NALDV-like genomes belonging to Classes I and II (Supplementary Fig. 5), indicating integration events happened very recently and maybe still happening in spiders. The sequence of nrEVEs

https://doi.org/10.1038/s41467-025-61035-2

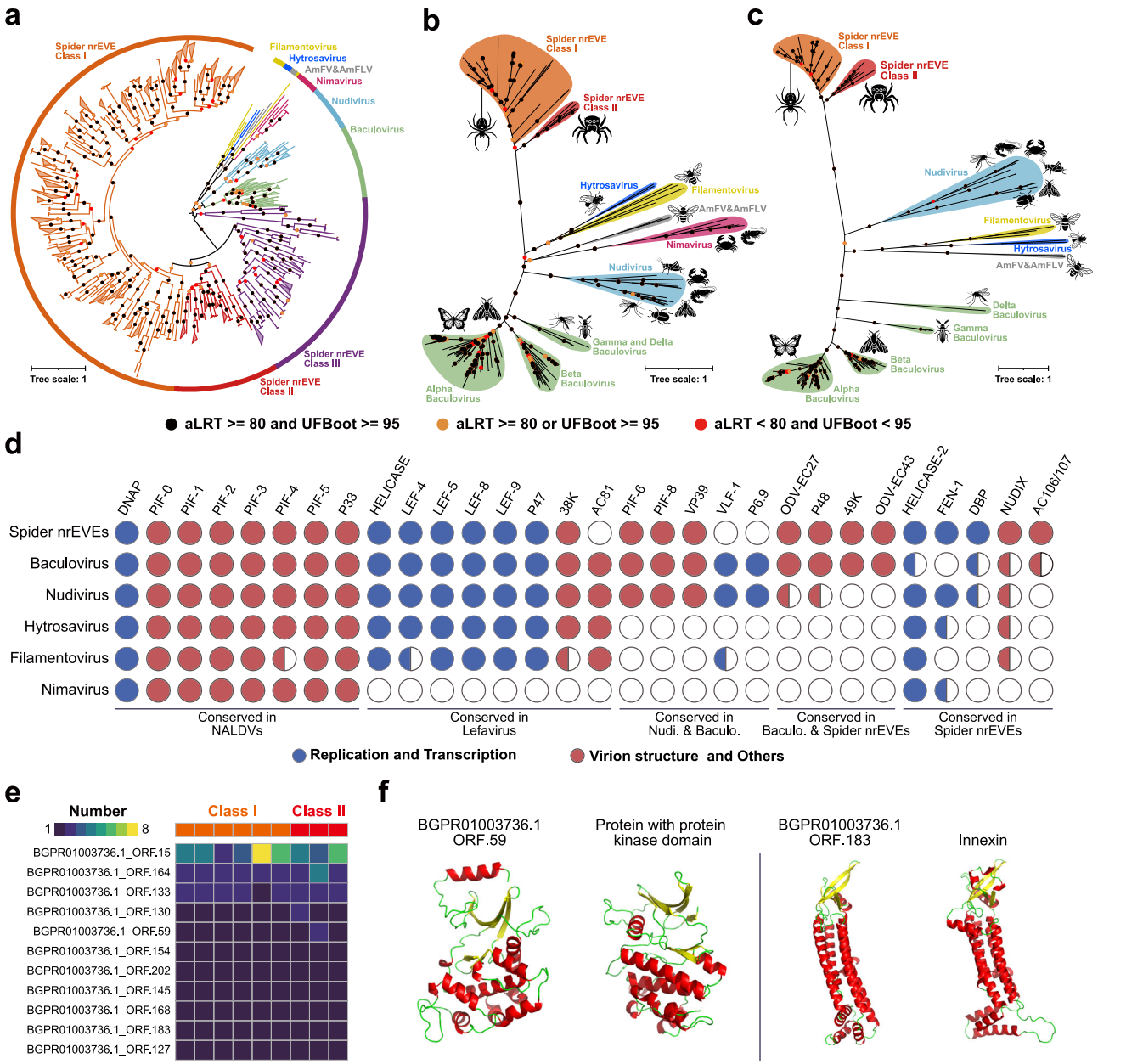


Fig. 5 | Phylogenetic analysis and gene consistent of spider nrEVEs. a A maximum-likelihood phylogenetic tree built from PIF-0 sequences (478 amino acid positions, LG + I + R10 model). Branches with an average branch length distance to their leaves below 0.5 were auto-collapsed. Several branches with abnormal length were pruned. **b** A maximum-likelihood phylogenetic trees built from concatenated PIFs (PIF-0, PIF-1, PIF-2, PIF-3, PIF-4, and PIF-5, 2233 amino acid positions, LG + I + R8 model). **c** A maximum-likelihood phylogenetic tree built from concatenated LEFs (LEF-9, P47, LEF-8, LEF-5, and LEF-4, 2407 amino acid positions) using the Q.yeast + F + I + R8 model. Values at the nodes in (**a**–**c**) indicate the branch supports calculated by the Shimodaira–Hasegawa-like approximate likelihood ratio test (aLRT) and ultrafast bootstrap approximation (UFBoot) methods with 1000 replicates. **d** The conserved gene composition between NALDVs and spider nrEVEs. Proteins are classed into two group based on their function. The solid circles indicate the protein is conserved in all the viruses of this group, semicircles indicate that the protein presence in some but not all of the viruses of the group, and the open circles indicate that the protein is absent. FEN-1 flap endonuclease 1, DBP DNA-binding protein, NUDIX nudix hydrolases; **e** Heatmap of the spider nrEVE-specific conserved proteins based on structure comparison. **f** Predicted structures of two conserved proteins in spider nrEVEs and their homologous structures in AFDB50. Proteins are colored based on secondary structure properties.

from different population of *A. ventricosus* showed variable patterns (Supplementary Fig. 3), suggesting the fixation of viral genes might still be ongoing at present.

Horizontal transfer and integration of viral genomes have been reported to play important roles in species evolution[45,46]; therefore, spider nrEVEs may play a role in the differentiation of spider species. The KaKs analysis of *A. ventricosus* indicated that spider nrEVEs were selected during evolution (Fig. 4c). Among NALDVs, integration events appeared to have happened at different time points during evolution. The presence of evolutionary distant baculoviral homologs and spider nrEVEs in same spider species suggests that the integration events occurred multiple times (Supplementary Fig. 5). The integration event of betanudivirus in Braconidae has been confirmed to be happened at ~100 Mya[35] and the integration of alphanudivirus occurred, at best, a few million years ago[29]. Given the large and widespread number of integration events in spiders, estimating the timing of specific viral integration events is very challenging. The phylogenetic analysis indicated that nimaviruses are the sister group of spider nrEVEs, as LEFs exist in all NALDVs other than nimaviruses, we assumed that the LEFs in nimaviruses might have been lost during evolution (Fig. 7).

https://doi.org/10.1038/s41467-025-61035-2

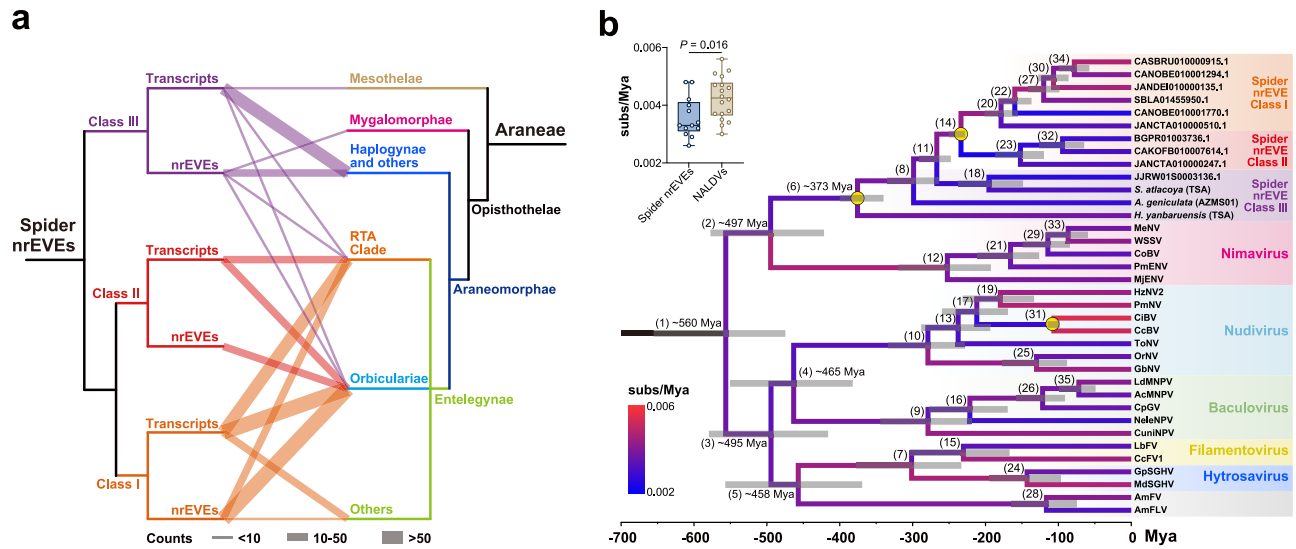


Fig. 6 | Relationship and molecular dating of spider nrEVEs and their spider hosts. a The relationship between the classifications of the identified spider nrEVEs and their spider hosts. The thickness of link lines represented the number of targets of PIF-0. **b** The evolutionary timeline of spider nrEVEs and NALDVs is illustrated. Calibrated nodes are denoted by circles, while semi-transparent bars represent the 95% highest posterior density (HPD) of age estimates for the main clades. Age estimates and their 95% HPD intervals for the corresponding nodes are provided in Table 2. The branches are colored by the substitutions per site per million years (subs/Mya). The boxplot in the upper left corner shows the substitution rates difference from spider nrEVEs ($n = 13$) and NALDVs ($n = 18$). Statistical analysis was performed using two-tailed Student's *t*-test. Boxplot boxes denote the interquartile range, middle lines indicate the median, and error bars indicate the min and max. The source data and *P* values were provided in Source Data file.

Collectively, these results confirmed that the common ancestor of NALDVs appeared ~560 Mya and that their integration events occurred continuously during the evolution of NALDVs and arthropods.

No NALDV nor NCLDVs have been identified in contemporary spiders. Whether exogenous viruses homologous to these nrEVEs exist in spiders is unclear. However, some nrEVEs in spiders maintain integrity of NALDV-like genome during evolution and under the pressure of purifying selection (Fig. 4c). Analysis of the genome composition of these nrEVEs revealed that, in addition to key viral genes for replication and transcription, they contain almost all the virion structure proteins conserved among various NALDVs, such as PIFs, VP39, 38K, 49K, ODV-EC27, and ODV-EC43[47]. It is tempting to speculate that these nrEVEs can potentially form complete viral particles. It is also possible that unidentified NALDVs with integration properties may be the cause by spider nrEVEs. The widespread discovery of a large number of viral mRNAs (Fig. 4a) and related sRNAs in spider (Fig. 4b) indicate that these viral remnants may still be active in spiders, or represent latent infection properties in spiders. This hypothesis needs further investigation and substantiation.

It remains unknown whether the nrEVEs have certain functions in spiders. mRNAs related to spider nrEVEs had higher expression patterns in the main silk glands (Fig. 4d), and the sRNA patterns in different parts of the main silk glands showed obvious preferences (Fig. 4b), indicating that spider nrEVEs may function in the silk gland. Spider nrEVEs were exclusively found in Araneae rather than other orders of arachnids in the present study, suggesting that spider nrEVEs may be one of the factors shaping the highly specialized and diverse silk glands of spiders. More in-depth functional and experimental studies, such as electron microscopy and protein level verification, may provide a clearer understanding of the role of spider nrEVEs.

In previous studies, the discovery of EVEs was primarily relied on sequence searches for known viral proteins, which were limited by sequence similarity[15,18,34]. Structural prediction and comparison play an important role in the discovery and classification of spider nrEVEs, suggesting that applying structural comparison could provide new insights for identifying and classifying distant viral sequences. Most recently, study based on predicted viral structural alignments have provided functional insights that were not achievable with previous approaches[48]. Additionally, our analysis showed that spider nrEVEs exhibited significant differences in ORF coverage and GC content compared to their adjacent sequences (Fig. 2c), aligning with previous findings that endogenous sequences of viruses often exhibit heterogeneity[31]. Genomic binning based on viral genome features has also recently been developed[30,49], and differences in the sequence characteristics between viral and eukaryotic genomes may serve as an effective method for identifying endogenous viral sequences.

We need to point out that our approach relies on the structural information of proteins. Currently, protein structures in the AFDB50 are predicted based on annotated protein sequences. There are many un-annotated proteins lack structural information in the database. Therefore, it is necessary to manually annotate those proteins and predict their structure for subsequent searches. It should also be noted that reasonable thresholds need to be set to filter out structures with low similarity. All structures used for subsequent analysis in this study were screened with probability of homologous over 0.9 to ensure that these structures were not false positives. Distant sequences discovered by structural alignments may require further analysis and verification to ensure that they are not the result of convergent evolution in structure. During our study, many distinct proteins from bacteria and metagenomic samples were also identified sharing high structure similarities with baculovirus proteins including DNAP, 38K, LEF-8, and Helicase (Supplementary Data 1), however, PCA revealed they were not grouped with species carrying NALDV-related nrEVEs (Fig. 1b), and were not further analyzed.

Summarily, in this study, the widespread presence of nrEVEs in spider genomes is identified through structural comparisons and sequence analyses, highlighting multiple integration events of unidentified NALDVs happened both anciently and recently in spiders. The discovery of spider nrEVEs has significantly expanded our understanding of nrEVEs in arthropods, as well as the host distribution and evolutionary history of NALDVs within these organisms. The integration of nrEVEs into the spider genome may have influenced spider evolution, more specifically in spider silk glands. These findings offer crucial insights into the specific functions, evolutionary

Table 2 | Age estimates and their 95% HPD intervals of spider nrEVEs and NALDVs

Node	Time, Mya	95% HPD, Mya	Prior	Classification
1	559.60	474.64–655.48		NALDVs
2	496.74	421.55–577.38		
3	494.69	415.93–579.66		Lefavirus
4	465.06	382.19–550.56		
5	458.45	369.22–557.11		
6	373.12	339.84–400.00	299–400	Araneae Spider nrEVEs
7	302.71	233.10–377.88		
8	299.10	264.54–335.53		
9	280.29	219.43–344.48		Baculovirus
10	280.09	227.94–333.73		Nudivirus
11	268.16	247.52–291.06		
12	254.20	192.59–319.70		Nimavirus
13	238.48	193.39–288.33		
14	235.78	228.00–250.32	228–386	Entelegynae Spider nrEVEs Classes I&II
15	231.44	166.96–303.66		Filamentovirus
16	222.51	169.44–277.40		
17	213.01	169.20–259.87		
18	194.55	148.39–237.97		
19	180.73	133.28–231.49		
20	178.53	155.41–201.37		Spider nrEVEs Class I
21	166.83	126.44–211.80		
22	159.62	136.85–182.21		
23	152.41	119.83–186.98		Spider nrEVEs Class II
24	143.96	96.80–194.34		Hytrosavirus
25	131.01	88.11–177.00		
26	122.33	90.85–157.70		
27	120.13	98.45–141.50		
28	117.93	75.29–165.13		AmFV & AmFLV
29	115.24	84.15–149.03		
30	106.52	85.96–127.78		
31	106.21	97.93–114.87	103.38 ± 4.41[35]	Bracovirus
32	94.90	64.52–126.77		
33	87.49	59.41–117.18		
34	78.32	57.53–99.55		
35	72.84	48.91–98.82		

mechanisms, and development of spider nrEVEs and silk glands. Furthermore, our study demonstrates the potential of protein structure prediction and comparison, alongside genomic features, for virus classification and the discovery of endogenous viruses.

Methods

Protein structure prediction and homologs search

Coding sequences of representative baculovirus species (Supplementary Data 9) were extracted from the NCBI viral genomic database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/; accessed April 2023). Protein structure models were predicted using LocalColabFold[50] (https://github.com/YoshitakaMo/localcolabfold) with default setting, and the top-ranked structure for each sequence was selected for subsequent homologous structure searches. The AFDB50 database was obtained from https://foldseek.steinegerlab.workers.dev on June 23, 2023. Homologous structure searches in AFDB50 were conducted using Foldseek[51] (v6.29e2557) with default mode (https://github.com/steineggerlab/foldseek), and the homologous cutoff value was set at 'prob' = 0.9. PCA and clustering was carried out on different taxon within AFDB50 harboring baculoviral core protein homologs using custom R script. The baculoviral homologous proteins identified in *A. ventricosus* were queried in the UniProt database, and the corresponding target scaffolds were downloaded from the NCBI WGS database.

All ORFs of the selected scaffolds in *A. ventricosus* and spider nrEVEs (Supplementary Data 9) were extracted using orfipy[52] (v0.0.3) with parameters "–min 150 –start ATG –ignore-case." Subsequently, the structures of these ORFs and RefSeqs of the NALDV proteins were predicted using LocalColabFold with the default setup. The predicted structures were used for "all-against-all" searches using Foldseek. The homologs were conducted through structure alignment results ('prob' > 0.9) using cytoscape[53], and confirmed via sequence alignments utilizing MSAProbs[54] (https://toolkit.tuebingen.mpg.de/tools/msaprobs).

Survey of baculoviral homologs in NCBI WGS and TSA database

The WGS data for Arachnida and TSA data for Araneae were obtained from NCBI in June 2023 (Supplementary Data 3 and 7). Chromosomes of *T. clavate* were downloaded from SpiderDB 1.0 (https://spider.bioinfotoolkits.net). All ORFs in the assemble sequences were extracted using orfipy with the parameters "–min 150 and –start ATG –ignore-case." Subsequently, the database of all peptide was built for homologous search using MMseqs2[55] (v14.7e284). The sequence of baculoviral homolog proteins identified by structure comparison of selected scaffolds in *A. ventricosus* were used as queries with default settings plus "–format-output query, target, evalue, bits, qseq, tseq, qcov."

Maximum likelihood phylogenetic analysis

For the phylogenetic analysis of single or concatenated PIFs and LEFs, sequences were first extracted from the MMseqs2 search results using custom R script and aligned using MAFFT[56] (v7.505). Aligned sequences were trimmed using TrimAl[57] (v1.4.1) with -gt 0.5 or -gt 0.8, and used for phylogenetic analyses using IQ-TREE[58] (v2.2.5) with ModelFinder Plus to determine the best-fit model. Support was computed from 1000 replicates for Shimodaira–Hasegawa (SH)-like aLRT and UFBoot. As per the IQ-TREE manual, supports were deemed good when SH-like aLRT ≥ 80% and UFBoot ≥ 95%.

Molecular dating

The sequences of concatenated PIF-0 and LEF-8 were used for molecular dating analysis. The divergence times were estimated using the Bayesian inference approach implemented in BEAST (v2.7.6)[59]. The divergence times of CiBV and CcBV[36], and the original fossil data of Mesothelae[38,39] and extant araneomorphs[37] were used as calibration points[35,40] (Table 2). We assumed uncorrelated lognormal relaxed clock and free mean substitution rates. The phylogeny was calibrated under a normal distribution and Yule process[35]. Monte Carlo Markov chains were run for 100 million generations, sampling every 1000 generation, using a burn-in of 1000. Age estimates and substitution rates were analyzed using the FigTree (v1.4.4).

DNA and RNA extraction of spiders

The wild spiders (*A. ventricosus* and *T. clavata*) captured in Nanyang, Henan Province, China, were purchased via Taobao (https://shop104568564.taobao.com). The collection of wild spiders complies with all regulations. For the DNA extraction of *A. ventricosus*, the spiders were first immersed in liquid nitrogen and macerated prior to DNA extraction. Total DNA extraction was carried out using the E.Z.N.A.® Insect DNA Kit (Omega Bio-tek, Norcross, USA) following the manufacturer's protocol. For RNA extraction from

https://doi.org/10.1038/s41467-025-61035-2

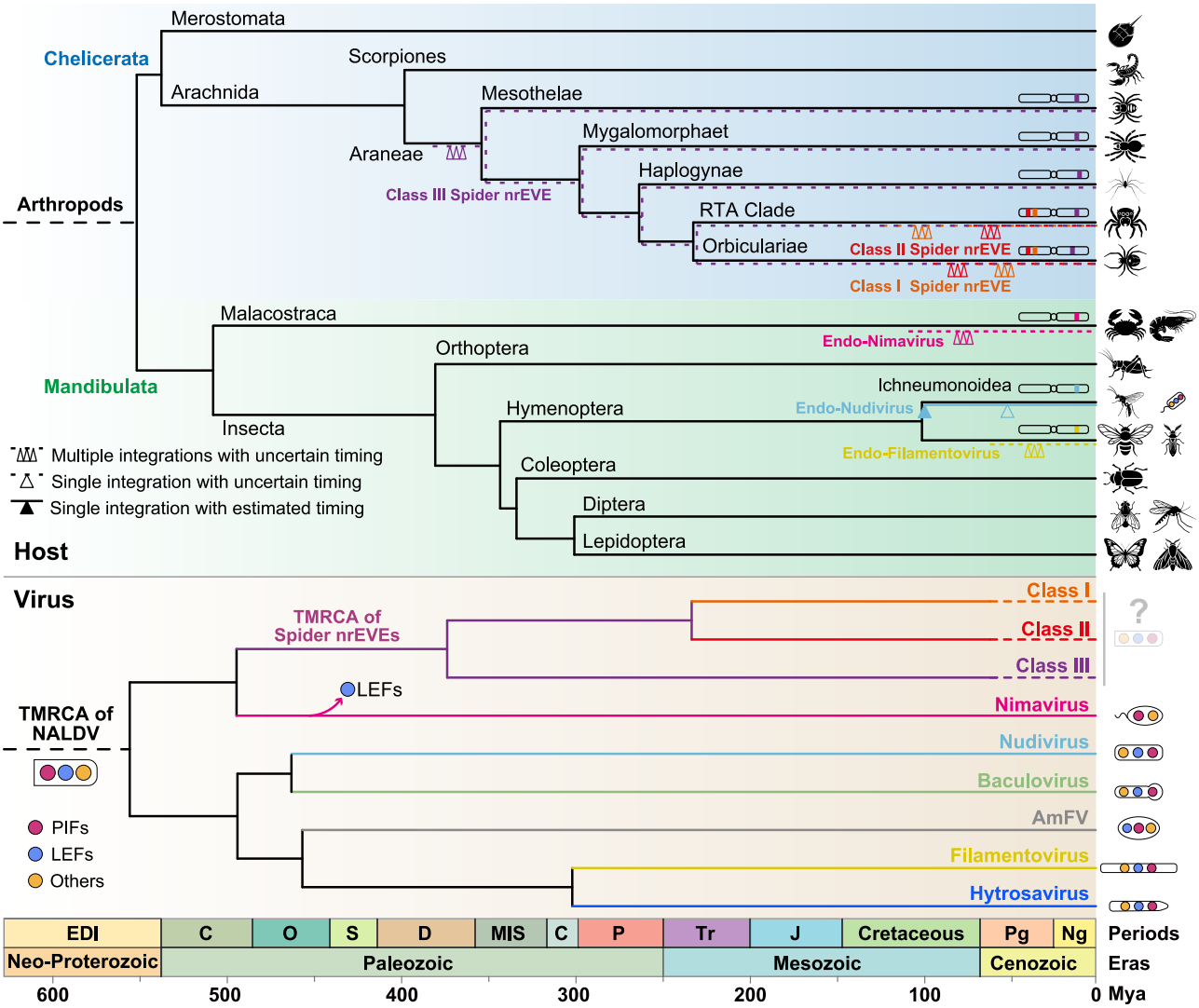


Fig. 7 | Long-term co-evolution and integration events between NALDVs and arthropods. Schematic depicting the integration and co-evolution history of arthropods and NALDVs. Multiple integration events of unidentified NALDVs occurred during the evolution of spiders. Ancient unidentified NALDVs related to class III nrEVEs were integrated into the TMRCA of spiders. These nrEVEs engaged in a long-term integration history with spiders and NALDVs related to classes I and II nrEVEs integrated into spiders more recently. The divergence time of NALDVs and their relevant nrEVEs or endoviruses were inferred from molecular dating analyses, and the host divergence times were based on Timetree[44]. The triangles and lines represent the integration of different viruses. Each virus branch and its corresponding integration events are color-coded according to the text labels. Different types of viral genes are symbolized by colored circles, with arrow on the evolutionary tree indicating the loss of these genes. The diagram above host branches shows the integration of different types of virus sequences on the chromosome. The diagram on the right side of the viral branches shows the morphology of different types of viruses and the types of viral genes they contain. The diagram of the bracovirus is shown next to Ichneumonoidea.

T. clavata, four fresh tissues (legs, cephalothorax, silk glands, and abdomen without silk glands) were dissected in PBS. The tissues were macerated using a tissue shredding tube with TRIzol reagent (Invitrogen), and total RNA was extracted following the manufacturer's protocol.

DNA sequence and data analysis

DNA from an adult *A. ventricosus* was used for library construction and shotgun sequencing at The Beijing Genomics Institute. The sequence data were aligned to the reference assembly of *A. ventricosus* (NCBI WGS accession number: BGPR01) by Bowtie2 (v2.5.3)[60] with default settings, and the consensus sequences of the scaffolds were extracted using samtools[61] (v1.20) mpileup function setting '-A -d 1000 -B -Q 0' and iVar[62] (v1.4.2) consensus function, which selects the most frequent base as the consensus base. Uncovered sites were replaced by 'N'. Ka/Ks analyses were conducted between the consensus sequences generated by our sequencing data and the reference assemblies of AvEVE1-9 via TB-Tools (v2.085)[63]. Briefly, the ORFs from our consensus sequences and the reference assemblies of AvEVE1-9 were extracted using orfipy with the parameters "–min 150 and –start ATG –ignore-case". Subsequent homology identification and gene integrity analysis were performed by screening the MMseqs2 results. For identified homologs, the amino acid and corresponding cDNA sequences of homologs within AvEVE1-9 conducted using the integrated 'Simple Ka/Ks Calculator' module in TB-Tools. Finally, the Ka/Ks values of the intact genes across identical AvEVEs from different populations were evaluated. The sRNA data of major silk glands of *T. clavata* were downloaded from the National Genomics Data Center (accession number: CRR644805-CRR644813), after mapping to the nuclear sequence of baculoviral homologs in *T. clavata* by 'Bowtie' with a perfect match, the analysis to follow were conducted by custom R script.

PCR and qRT-PCR

For PCR assay, 2 μL spider genome DNA was used as template and PCR was conducted by Phanta Max Master Mix (Vazyme). The amplicons were electrophoresed on 0.8% agarose gel. For reverse transcription experiments, 0.5 μg total RNA was used to synthesize cDNA by HiScript III RT SuperMix (Vazyme). cDNA was used for quantitative PCR assay using the StepOne Plus Real-time PCR system (Applied Biosystems) with TB Green Premix Ex Taq II (Takara). The primers used for PCR and qPCR are listed in Supplementary Data 12.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Databases used in our study include: NCBI RefSeq (https://ftp.ncbi.nlm.nih.gov/refseq/); NCBI WGS (Whole Genome Shotgun) genomes (https://ftp.ncbi.nlm.nih.gov/genbank/wgs/); NCBI TSA (Transcriptome Shotgun Assemblies) (https://ftp.ncbi.nlm.nih.gov/genbank/tsa/); UniRef90 (https://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref90/); National Genomics Data Center (https://ngdc.cncb.ac.cn/); AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/); and SpiderDB 1.0 (https://spider.bioinfotoolkits.net). Accession codes for all sequence data analyzed in this study were provided in Supplementary Data files. The high-throughput sequencing raw data of *A. ventricosus* genomic DNA were deposited into the Genome Sequence Archive (GSA) of the National Genomics Data Center and are available through accession number: CRA017441. The original data generated in our study, including sequence alignments, phylogenetic trees, predicted protein structures, etc., have been made publicly available at Figshare (https://doi.org/10.6084/m9.figshare.26860672). Source data are provided with this paper.

Code availability

Custom scripts for data analysis are available at Figshare (https://doi.org/10.6084/m9.figshare.26860672).

References

1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).
2. Holmes, E. C. The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368–377 (2011).
3. Weiss, R. A. The discovery of endogenous retroviruses. *Retrovirology* **3**, 67 (2006).
4. Horie, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
5. Houe, V., Bonizzoni, M. & Failloux, A. B. Endogenous non-retroviral elements in genomes of Aedes mosquitoes and vector competence. *Emerg. Microbes Infect.* **8**, 542–555 (2019).
6. Huang, H.-J. et al. Co-option of a non-retroviral endogenous viral element in planthoppers. *Nat. Commun.* **14** https://doi.org/10.1038/s41467-023-43186-2 (2023).
7. Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* **25**, 81–89 (2017).
8. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: connecting recent and ancient viral evolution. *Virology* **479-480**, 26–37 (2015).
9. Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* **17**, 355–370 (2019).
10. Regier, J. C. et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083 (2010).
11. van Oers, M. M. et al. Developments in the classification and nomenclature of arthropod-infecting large DNA viruses that contain pif genes. *Arch. Virol.* **168**, 182 (2023).
12. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. R.* **84**, e00061–19 (2020).
13. Passarelli, A. L. & Guarino, L. A. Baculovirus late and very late gene regulation. *Curr. Drug Targets* **8**, 1103–1115 (2007).
14. Lin, C. L. et al. Persistent Hz-1 virus infection in insect cells: evidence for insertion of viral DNA into host chromosomes and viral infection in a latent status. *J. Virol.* **73**, 128–139 (1999).
15. Cheng, R. L., Li, X. F. & Zhang, C. X. Nudivirus remnants in the genomes of arthropods. *Genome Biol. Evol.* **12**, 578–588 (2020).
16. Bezier, A. et al. Polydnaviruses of braconid wasps derive from an ancestral nudivirus. *Science* **323**, 926–930 (2009).
17. Strand, M. R. & Burke, G. R. Polydnaviruses: evolution and function. *Curr. Issues Mol. Biol.* **34**, 163–181 (2019).
18. Kawato, S. et al. Crustacean genome exploration reveals the evolutionary origin of white spot syndrome virus. *J. Virol.* **93** https://doi.org/10.1128/JVI.01144-18 (2019).
19. Guinet, B. et al. Endoparasitoid lifestyle promotes endogenization and domestication of dsDNA viruses. *Elife* **12** https://doi.org/10.7554/eLife.85993 (2023).
20. Legg, D. A., Sutton, M. D. & Edgecombe, G. D. Arthropod fossil data increase congruence of morphological and molecular phylogenies. *Nat. Commun.* **4**, 2485 (2013).
21. Kulkarni, S., Wood, H. M. & Hormiga, G. Advances in the reconstruction of the spider tree of life: a roadmap for spider systematics and comparative studies. *Cladistics* **39**, 479–532 (2023).
22. Sanggaard, K. W. et al. Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Commun.* **5**, 3765 (2014).
23. Mihara, T. et al. Linking virus genomes with host taxonomy. *Viruses* **8**, 66 (2016).
24. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01773-0 (2023).
26. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
27. Liu, S., Coates, B. S. & Bonning, B. C. Endogenous viral elements integrated into the genome of the soybean aphid, Aphis glycines. *Insect Biochem. Mol. Biol.* **123**, 103405 (2020).
28. Burke, G. R., Simmonds, T. J., Sharanowski, B. J. & Geib, S. M. Rapid viral symbiogenesis via changes in parasitoid wasp genome architecture. *Mol. Biol. Evol.* **35**, 2463–2474 (2018).
29. Pichon, A. et al. Recurrent DNA virus domestication leading to different parasite virulence strategies. *Sci. Adv.* **1**, e1501150 (2015).
30. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
31. Hackl, T., Duponchel, S., Barenhoff, K., Weinmann, A. & Fischer, M. G. Virophages and retrotransposons colonize the genomes of a heterotrophic flagellate. Elife 10, e72674 (2021).
32. Li, W. H. *Molecular Evolution* (Sinauer Associates, 2007).
33. Guinet, B. et al. A novel and diverse family of filamentous DNA viruses associated with parasitic wasps. *Virus Evol.* **10**, veae022 (2024).
34. Ter Horst, A. M., Nigg, J. C., Dekker, F. M. & Falk, B. W. Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs. *J. Virol.* **93**. https://doi.org/10.1128/JVI.02124-18 (2019)
35. Thézé, J., Bézier, A., Periquet, G., Drezen, J.-M. & Herniou, E. A. Paleozoic origin of insect large dsDNA viruses. *Proc. Natl. Acad. Sci. USA* **108**, 15931–15935 (2011).

36. Murphy, N., Banks, J. C., Whitfield, J. B. & Austin, A. D. Phylogeny of the parasitic microgastroid subfamilies (Hymenoptera: Braconidae) based on sequence data from seven genes, with an improved time estimate of the origin of the lineage. *Mol. Phylogenet. Evol.* **47**, 378–395 (2008).
37. Selden, P. A., Anderson, H. M. & Anderson, J. M. A review of the fossil record of spiders (Araneae) with special reference to Africa, and description of a new specimen from the Triassic Molteno Formation of South Africa. *Afr. Invertebrates 50* **1**, 105 (2009.
38. Selden, P. A. Fossil mesothele spiders. *Nature* **379**, 498–499 (1996).
39. Selden, P. A., Shear, W. A. & Bonamo, P. M. A spider and other arachnids from the Devonian of New York, and reinterpretations of Devonian Araneae. *Palaeontology* **34**, 241–281 (1991).
40. Bond, J. E. et al. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Curr. Biol.* **24**, 1765–1771 (2014).
41. Wallau, G. L. RNA virus EVEs in insect genomes. *Curr. Opin. Insect Sci.* **49**, 42–47 (2022).
42. Frank, J. A. et al. Evolution and antiviral activity of a human protein of retroviral origin. *Science* **378**, 422–428 (2022).
43. Gilbert, C. & Belliardo, C. The diversity of endogenous viral elements in insects. *Curr. Opin. Insect Sci.* **49**, 48–55 (2022).
44. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
45. Arnold, B. J., Huang, I. T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol* **20**, 206–218 (2022).
46. Gilbert, C. & Cordaux, R. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr. Opin. Virol.* **25**, 16–22 (2017).
47. Jia, X. et al. Architecture of the baculovirus nucleocapsid revealed by cryo-EM. *Nat. Commun.* **14**, 7481 (2023).
48. Nomburg, J. et al. Birth of protein folds and functions in the virome. *Nature* **633**, 710–717 (2024).
49. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res.* **50**, e83 (2022).
50. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
51. Van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
52. Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
53. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
54. Gabler, F. et al. Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinforma.* **72**, e108 (2020).
55. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
56. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
57. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
58. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
59. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10** https://doi.org/10.1093/gigascience/giab008 (2021).
62. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
63. Chen, C. et al. TBtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant* **16**, 1733–1742 (2023).

Acknowledgements

This study was supported by the grants from the National Natural Science Foundation of China (No. 32400133 to H.H., 32370182 to M.W., and 32170156 to Z.H.). We thank all the databases and projects that provide the public sequencing data for this study.

Author contributions

H.H., Z.H., and M.W. conceptualized and designed the study. H.H. performed all computational analyses and experiments. H.H., Z.H., M.W., and J.M.V. interpreted the results and wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61035-2.

Correspondence and requests for materials should be addressed to Zhihong Hu or Manli Wang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.