Published on Aug 09, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-14b2-6f27

Identification of capsid-like proteins in venomous and parasitic animals

Inspired by wasps co-opting viral capsids to deliver genes to the caterpillars they parasitize, we looked for capsid-like proteins in other species. We found capsid homologs in ticks and other parasites, suggesting this phenomenon could be more widespread than previously known.

Contributors (A-Z)

Audrey Bell, Brae M. Bigge, Adair L. Borges, Seemay Chou, Rachel J. Dutton, Jase Gehring, Megan L. Hochstrasser, Elizabeth A. McDaniel, Kira E. Poskanzer, Taylor Reiter, Ryan York

Version 4 · Mar 31, 2025

Purpose

The development of AAV capsids for therapeutic gene delivery has exploded in popularity over the past few years. However, humans aren't the first or only species using viral capsids for gene delivery — wasps evolved this tactic over 100 million years ago. Parasitoid wasps that lay eggs inside arthropod hosts have co-opted ancient viruses for gene delivery to manipulate multiple aspects of the host's biology, thereby increasing the probability of survival of the wasp larvae **[1][2]**.

We wondered if venomous species that bite humans have also evolved to use viral capsids to deliver molecules that manipulate their hosts. We used a multi-pronged sequence- and structure-based search for viral capsids across venomous species and found evidence for endogenized viral capsids, most notably in ticks. Though we cast a broad net in this effort, we were most interested in finding novel nucleic acid delivery systems in species that deliver cargo to humans, as these would be the most useful therapeutic modalities. Here, we focused on finding endogenized capsid proteins, and our follow-up work has focused on developing methods to specifically identify potential nucleic acid cargos in parasites **[3]**.

These findings may be useful to other scientists interested in the domestication of viral genes, especially in the context of parasitism.

- **Data** from our capsid HMM search and our preHGT run to find putatively endogenized viral capsids are available on Zenodo, as is our follow-up work looking into individual hits with BLAST searches of tick salivary transcriptome, genomic neighborhood analysis, and Foldseek analysis.
- **Data** from our ProteinCartography runs, including structures of all tick and viral capsid proteins, the configuration file, and all outputs, are available on Zenodo. We performed ProteinCartography analysis using version 0.4.2 of the pipeline, found in this GitHub repository.
- All associated **code** is available in this GitHub repository.

We’ve put this effort on ice! ☒

#DeadEnd

We found intriguing capsid-like proteins across several parasitic species. To advance, we'd need to figure out what cargo the putative capsids deliver, if any. In a follow-up study, we tried to computationally identify packaged nucleic acids in parasites, as we’re most interested in therapeutic nucleic acid delivery, but our results weren’t sufficiently compelling to warrant further investment **[3]**.

Learn more about the Icebox and the different reasons we ice projects.

Background and goals

Capsid-based delivery of therapeutics represents a growing modality of interest across the biotech/pharma industry. The focus of most recent efforts has been on a specific class of capsids from adenovirus-associated viruses (AAVs). AAV-based therapies have been approved for use in humans, and researchers are engineering and applying these vectors for a variety of indications **[4]**. However, co-opting viral capsids for gene delivery is an ancient innovation, dating back over 100 million years.

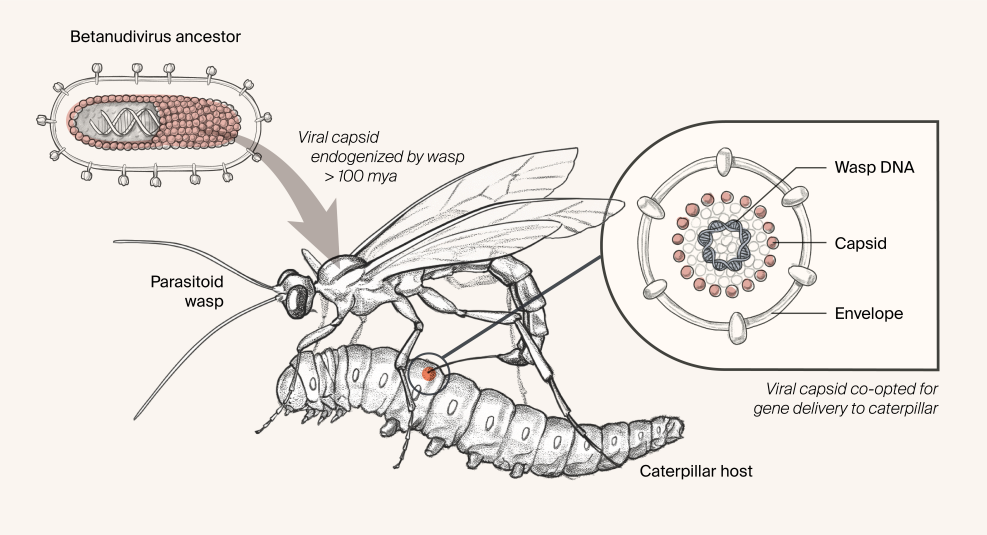

Figure 1

Braconid parasitoid wasps use domesticated viruses to protect their eggs from the host immune system.

Braconid parasitoid wasps like *Microplitis demolitor* lay their eggs in living lepidopteran hosts, where the larvae consume the host as a food source throughout development. To protect their eggs from the caterpillar immune response, the wasps evolved a fascinating strategy of deploying an ancient endogenized betanudivirus (called a bracovirus) as a gene delivery agent. The wasps inject viral particles containing wasp DNA alongside their eggs, and prolonged expression of these genes in the caterpillar alters the host immune system to protect the developing larvae.

mya: Million years ago

Scientists uncovered this innovation by studying parasitoid wasps, which use viral capsids derived from ancient polydnaviruses to deliver genes to their caterpillar hosts (Figure 1) **[5][6]**. While there is an enormous diversity of ecologies and lifestyles across parasitoid wasps, many parasitize other arthropods, especially lepidopterans. These wasps inject their eggs into caterpillars, and the developing larvae use the host as a source of food. To help protect the eggs, the parasitoid wasps co-inject viral-like particles that manipulate the caterpillar's biology, including its immune system **[7]**. The injected packages are made up of viral capsids filled with wasp DNA that helps to protect the wasps's larvae. Once injected into the host, wasp DNA is integrated into

the host genome to drive expression of host-manipulating factors, such as immune-dampening proteins.

Wasps have co-opted distinct types of polydnaviruses multiple times **[8]**, with different wasp lineages harboring distinct endogenized viruses that they deploy for host manipulation. Of these, some of the best-studied are the bracoviruses found in the genomes of braconid wasps like *Microplitis* and *Cotesia.* Although the bracovirus capsid proteins play a key role in delivery of nucleic acid cargo to host species, studies on bracovirus biology show key roles for other virus-derived machinery in the process — especially packaging machinery and envelope proteins.

Our goal in this pilot was to explore whether other venomous species have similarly evolved the ability to use viral capsids for molecular delivery. Though we searched for endogenized capsids of any sort, we'd hoped to identify capsids that deliver genes to a human host. If we can find such capsids, they might serve as new starting points for developing gene delivery systems.

We used a combination of computational approaches to scan for the presence of capsid-like proteins across venoms, including sequence-, HGT-, and structure-based searches. We ultimately found capsid-like genes in the venom-like saliva of ticks as well as in the genomes of other parasitic species, including parasitic nematodes (helminths), mosquitoes, and aphids. Jump straight to these results or read on to learn the details of our approach.

The approach

To identify capsid-like proteins in venomous species, we applied two strategies: a **sequence homology-based search** for capsids in venom gland transcriptomes (HMM-based detection) and our **horizontal gene transfer (HGT) detection pipeline** to detect genes of putative viral origin in genomes of venomous species (preHGT). We then followed up on the most promising hits by searching for structural homologs across the AlphaFold database using Foldseek. Last, we experimented with a structural clustering-based approach to identify capsid-like structures in salivary transcriptomes but didn't pursue this approach extensively as it wasn't scalable and results were hard to interpret.

An HMM-based search of venom transcriptomes for viral capsids

We first performed a sequence-based search for capsid-like proteins in venom transcriptomes using hidden Markov models (HMMs) of viral capsids. A curated database of viral protein families and associated HMM profiles for each family is available for download from the viral orthologous groups (VOGs) database (version vog217). In addition, because bracovirus capsids aren't included in the VOG database, we also generated a custom HMM profile to include in searches based on a PSI-BLAST search results using the *Microplitis demolitor* major bracovirus capsid protein vp39 as a seed sequence.

We then used these HMMs to search against 156 transcriptomes from 135 species using HMMER3 **[9]**. To maximize our ability to find distant homologs, we initially returned all hits from the HMMER search with no E-value cutoff and followed up with manual filtering steps.

> You can find the **list of species** included in this database, along with all **scripts, metadata, and results**, in this GitHub repository.

Filtering results from the HMM-based search

We first filtered results to include only hits from a subset of VOGs containing the word “capsid,” but not “capsid maturation protease” or “capsid assembly protease,” in their annotation, as we found these VOGs don’t encode actual capsid proteins themselves, and were driving a lot of spurious hits to animal proteases. Next, we manually curated the output file in an attempt to further filter out false positives. This analysis suffered from three major sources of false positives: hits from non-endogenized viruses that are part of the venom virome, hits to venom proteins that don't appear to be endogenized capsids but share a common domain with the capsid HMM (for instance, a protease domain or an immunoglobulin domain), and spurious hits to the animal proteome due the fact we didn't use any E-value cutoff for our initial HMMER search.

To identify false positives, we performed BLASTp on the top 210 hits from the HMMER3 search (E-value 1.20×10^{-81}–2.60×10^{-3}). We performed these searches using the NCBI web server in March of 2023. If the protein had close hits to viruses but no

eukaryote genomes, we decided that it could be a contaminant virus and not endogenized. If the protein had hits to proteins from other eukaryotic genomes but was consistently annotated to encode a non-capsid protein, we considered it to be a false positive. Some predicted proteins repeatedly came up as false-positive hits, which likely reflects a shared domain with our viral HMMs. Common hits that fell into this category were Clp protease, ovarian tumor domain (OTU), LRR domain, IG domain, hemicentin, obscurin, peroxidasin, titin, twitchin, fasciclin, neural cell adhesion molecule, roundabout, AAA ATPase, and 26S proteasome subunit. After manual curation, we were left with 14 putative endogenized capsids (not counting our expected, positive-control hit — the vp39 HMM to the *Microplitis demolitor* major bracovirus capsid protein vp39). The highest-confidence endogenized capsids we recovered were from four parasitoid wasp species and one spider.

> Download our **manually annotated results** from searching VOG HMMs against venom transcriptomes on Zenodo.

Horizontal gene transfer (HGT) analysis of venomous animal genomes

Because the evolutionary acquisition of viral genes is a key example of inter-kingdom horizontal gene transfer (HGT), we decided to also deploy an approach that was specifically tailored to detect inter-kingdom HGT events. We used our preHGT **[10]** pipeline to scan genus-level pangenomes for recent viral HGT events using a BLAST-based taxonomic approach. We used gene models for nine venomous species: *Ixodes scapularis, Ixodes persulcatus, Microplitis demolitor, Sepia pharaonis, Trichomalopsis sarcophagae, Ampulex compressa, Ophiophagus hannah, Bothrops jararaca,* and *Naja naja*.

We created nucleotide pangenomes at the genus level by clustering genes at 90% length and sequence identity to determine the unique set of genes for the genus. We then chose a representative sequence for each cluster by selecting the sequence with the most alignments. We used BLASTp to compare representative proteins to a clustered non-redundant (nr) database, inspired by NCBI's ClusteredNR. The clustered database **[11]** allowed us to capture more taxonomically diverse hits.

We then used the “alien index” to predict whether an HGT event occurred **[12]**. The alien index is a metric that tells us how similar a given gene is to genes in related, non-self “acceptor” species, compared to unrelated potential “donor” species. In the case of horizontal gene transfer, we should see genes that have high similarity to genes in the donor species and low similarity to genes in species closely related to the acceptor species. We calculate the alien index by subtracting the E-value of the best donor (viral) hit from the E-value of the best non-self acceptor hit. If the best viral hit E-value is closer to zero than the best venomous species hit, the alien index will be positive. We used the following previously published thresholds **[13]** for determining HGT events: an alien index > 0 is possible HGT, > 15 is likely HGT, and > 45 is highly likely. We removed potential contaminants from this candidate list that we only saw in one genome within an investigated genus and that either had very high (> 90%) identity to a viral sequence or that were very short with high identity (< 20,000 bp, > 70% identity). We performed ortholog annotation (KEGG, PFAM, and viral and biosynthetic gene clusters) for proteins that may have been horizontally transferred.

The pipeline produced a table of possible HGT events, including the predicted donor and acceptor taxa, alien index and BLASTp values, ortholog annotations, and genomic location information for this protein in the “acceptor” genome. Each hit is scored based on its likelihood to be either a real HGT event vs. contamination. We filtered out all events that were scored as “likely contamination.” After filtering, we detected 66 putative HGT events, 21 of which were inter-kingdom events predicted to be from a viral donor.

For the six putative viral HGT-derived genes in ticks (*Ixodes persulcatus* and *Ixodes scapularis*), we used tBLASTn to see if similar genes were also integrated into the *Amblyomma americanum* genome, which we hadn’t included in this HGT analysis since we didn't have gene predictions for this species at the time of analysis. We observed multiple hits for KAG0427517.1 and KAG0420414.1 in the *A. americanum* genome, suggesting that such viral HGT events may be common in ticks.

> You can find our manually annotated results from running preHGT on the genomes of venomous species here and the genomes we input into preHGT, both on Zenodo.

> You can find the **preHGT GitHub repo** here.

Genomic context analysis of putative endogenized viral genes

To evaluate if viral genes are integrated into the eukaryotic genome (rather than a viral contaminant), we pulled the contigs for each hit from the DBSOURCE field on the protein's GenBank page. We displayed these contigs using Geneious Prime (version 2022.2.1) and manually analyzed the contigs to assess if they're part of the tick genome by looking for introns, a eukaryote-specific feature, in the capsid gene and predicting origin and function of neighboring genes using the BLAST, HHpred, and Foldseek web servers.

> You can find **GenBank files containing the genomic neighborhood** of putatively horizontally transferred viral genes in the tick genome here. We've shared seven GenBank files on Zenodo — one for each contig, named with the putative viral protein identifier.

Salivary expression analysis of putative endogenized viral genes

To evaluate whether our viral proteins of interest are expressed in tick saliva, we BLASTed our top seven candidate proteins against *Ixodes scapularis and Ixodes ricinus* proteins from the NCBI TSA (accession numbers: GADI01P.1, GEGO01P.1, GHXN01.1, GIFC01P.1, and GKHW01.1) using Geneious Prime (version 2022.2.1).

> Download our **BLAST results** from searching putatively endogenized viral proteins against tick salivary transcriptomes on Zenodo.

Foldseek analysis of putative endogenized viral genes

To search for structural homologs of our capsid proteins of interest in other species, we used Foldseek, a tool that enables structure-based searches for similar proteins **[14]**. We prioritized three of the hits from the HMM- and HGT-based search (*Ixodes* rhabdovirus nucleocapsid KAG0427517, *Ixodes* sabavirus nucleocapsid EEC17452, and *Ixodes* Gag-Pol XP_042148722). We also queried with the wasp bracovirus capsid NP_001401748 that originally inspired this project. There was a structure available in the AlphaFold database for the *Ixodes* sabavirus nucleocapsid (AF-B7QF30), but not for the other three proteins. We folded the other three using ColabFold (version 1.5.2) **[15]** and used these to search for structures with the Foldseek web tool (version Foldseek 5-53465f0).

We searched all three AlphaFold databases available on Foldseek (i.e., AlphaFold/UniProt50 version 4, AlphaFold/Swiss-Prot version 4, and AlphaFold/Proteome version 4) **[15][16]**. One important note is that the AlphaFold database specifically excluded any viral proteins, so any protein hits are likely not coming from the viruses themselves. We then filtered these hits using Foldseek E-values (similar to BLAST E-values), to determine the confidence of their hits. The developers of Foldseek consider E-values less than 0.01 to represent homologous pairs **[17]**, so we applied that cutoff here as well for the bracovirus capsid, the *Ixodes* sabavirus nucleocapsid, and the *Ixodes* rhabdovirus nucleocapsid, which each returned a low number of hits (< 20). However, the *Ixodes* Gag-Pol protein had a very wide distribution, so we used a more stringent E-value cutoff of 10^{-3} to narrow in on top hits, identifying 924 structural homologs.

> Download a CSV containing **results from searching putatively endogenized viral capsid proteins against AlphaFold**/UniProt50 (version 4), AlphaFold/Swiss-Prot (version 4), and AlphaFold/Proteome (version 4) using the Foldseek web server.

Structural clustering to identify capsids in *Ornithodoros turicata*

We tested a different approach to identify capsids that have been endogenized by ticks using structure-based clustering and the ProteinCartography pipeline **[18]**. We co-clustered viral proteins with tick proteins and identified clusters where tick proteins co-cluster with known viral capsids.

For this analysis, we first needed to obtain virus capsid protein structures for comparison. As stated previously, the AlphaFold database doesn't currently include viral proteins. Therefore, we performed structural predictions using ESMFold for all capsid-related proteins from the VOG database. We chose to use ESMFold in this case because it's faster than AlphaFold, which was important for the number of proteins we needed to fold **[19]**.

> These **predicted capsid protein structures** are on Zenodo in the "structures.zip" file.

We searched UniProt for all proteins from *Ornithodoros turicata* (UniProt organism ID: 34597). We chose to start with this particular tick because there are a tractable number (< 10,000) of structures from *O. turicata* on UniProt, and many of these structures were predicted from its salivary transcriptome. We then downloaded the associated metadata and structures using scripts from the ProteinCartography repo **[18]**.

> **Python scripts** for downloading metadata are here, the **script** for downloading structures is here, and **notebooks** to facilitate metadata and structure downloads are in this GitHub repo.

To test this co-clustering approach, we performed ProteinCartography clustering analyses on combinations of viral structures and tick structures. This run contained the entire collection of capsid-related VOG proteins and all of the tick proteins from *O. turicata*. ProteinCartography output files include interactive UMAP and t-SNE embeddings with metadata overlays. Also included in ProteinCartography's outputs are the corresponding TSV files that contain plotting information and all the metadata

for each analysis, an interactive heatmap showing the similarity within and between clusters, and all the intermediate outputs. We looked for structural clusters in which viral proteins co-clustered with tick proteins. We found that the *Ornithodoros* proteins A0A2R5L5R2 and A0A2R5L7H1 cluster with proteins from VOG00029, VOG05312, VOG06972, and VOG20608.

> The **data** we used for this analysis are on Zenodo, the **code** is in this GitHub repo, and **additional notebooks** to facilitate metadata and structure downloading and custom plotting are in this GitHub repo.

Taxonomic tree visualizations

We used the scientific species names from each analysis to create a taxonomic tree using phyloT (version 2) and the phyloT database (version 2023.2). Using the phyloT web server, we set the node identifiers to scientific names, used expanded internal nodes, set polytomy to yes, and exported the trees as Newick files that we visualized and further annotated in iTOL (version 6.9) **[20]**.

> **Newick files for the trees** in Figure 3 and Figure 5 are available on Zenodo.

> All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.12809305), including scripts for performing HMM and BLAST searches, plus notebooks for ProteinCartography analyses.

The results

SHOW ME THE DATA: Access data from our **capsid HMM search**, our **preHGT run** to find putatively endogenized viral capsids, and our **follow-up work** looking into individual hits with BLAST searches of tick salivary transcriptome, genomic neighborhood analysis, and Foldseek analysis on Zenodo (DOI: 10.5281/zenodo.12775362).

Data from our ProteinCartography runs, including structures of all tick and viral capsid proteins, the configuration file, and all outputs, are in a separate Zenodo record (DOI: 10.5281/zenodo.12796464).

Sequence-based searches highlight retrotransposon capsid-like genes in parasitoid wasps and other parasitic species, including ticks

We first used a sequence-similarity approach to search for capsid-like sequences in venomous species. Because there's a large body of public data collected from venom samples, we were able to curate a database of 156 venom transcriptomes from 135 different species **[21]**. We searched this database using viral capsid-specific hidden Markov model (HMM) profiles from the VOG database. Because the bracovirus capsid isn't represented in the VOG database, we also generated a custom bracovirus capsid HMM to use in searching against the venom transcriptomes.

Searches using VOG capsid HMMs generated over 3,000 hits before filtering. However, the bracovirus vp39 capsid HMM only hit the *Microplitis demolitor* transcriptome and didn’t have any other high-confidence hits. We conclude that bracovirus-like capsids are not broadly present across animal venoms. That said, this transcriptome-based search likely underrepresents the true distribution of domesticated bracovirus capsids within parasitoid wasps. Previous work on bracoviruses shows that they’re primarily present in the calyx fluid from wasp ovaries,

which is injected during oviposition along with venom gland secretions. While the *Microplitus demolitor* transcriptome we used came from both the wasp venom gland and ovaries, the other parasitoid data we used only included transcripts from venom glands **[5]**. The lack of ovary transcriptome data may, in part, explain the absence of bracovirus sequences from those parasitoid wasp transcriptomes.

We performed manual curation of the top 210 VOG HMM hits via BLAST, which suggested that many of the hits were false positives. Some false positives came from potential contaminating or venom virome-associated sequences, including bacteriophage capsid proteins. Other false positives came from hits of our viral HMMs to shared domains in eukaryotic genes that are probably not derived from recent viral capture and endogenization. Many of our false positive hits had immunoglobulin domains or protease domains, which are common to both viral proteins and host proteins.

These filters limited our homologs of interest to 14 total hits (not including the vp39 bracovirus capsid hit in *Microplitis demolitor)* (Figure 2). These viral hits came from four species of parasitoid wasp (*Cotesia vestalis*, *Diadromus collaris*, *Microplitis demolitor*, and *Pteromalus puparum*), and one spider (*Phoneutria nigriventer*). We were initially disappointed in the results from this search because it mostly returned endogenized proteins in wasps, which are already known to have domesticated capsids. Our goal had been to identify endogenized capsid proteins outside of parasitoids, ideally in species of venomous animals that bite humans or other mammals. However, when examining the taxonomic distribution of the 14 hits using BLAST, we found that one of these proteins — a Gag-Pol protein from LTR-type retrotransposons — had homologs outside of parasitoid wasps, including in ticks, which bite and feed on humans.

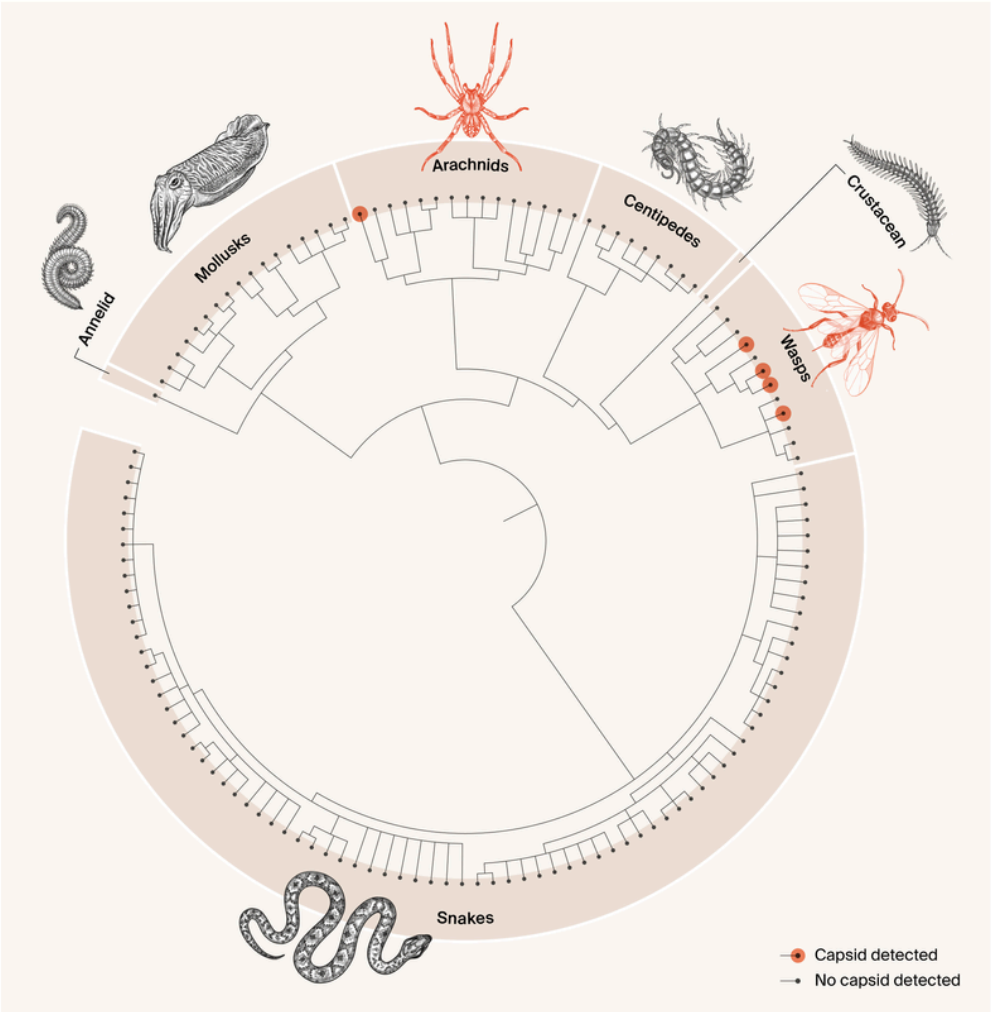

Figure 2

A sequence-based search of 135 species' transcriptomes identifies putatively endogenized viral capsids expressed in venom.

We searched venom transcriptomes from 135 species, which included venomous snakes, an annelid, mollusks, arachnids, centipedes, a crustacean, and parasitoid wasps. These species are displayed in a taxonomic tree. We did not find evidence of endogenized capsids in the venoms of most animals, indicated with black circles on leaf tips. However, four species of parasitoid wasp (*Cotesia vestalis*, *Diadromus collaris*, *Microplitis demolitor,* and *Pteromalus puparum*) have putatively endogenized capsids in their venoms, as does the Brazilian wandering spider (*Phoneutria nigriventer*). These instances of potential domestication are highlighted with red circles. Individual species names are omitted

from this figure due to the size and complexity of the tree, but the associated Newick file can be found here, the list of input transcriptomes is here, and the hits from this sequence-based search are here.

Interestingly, retrotransposon-derived Gag proteins form capsid-like structures, also known as viral-like particles (VLPs) **[22]**. Further, Gag-Pol homologs from humans (Peg10) and *Drosophila* (Arc) were recently shown to be involved in RNA delivery between cells **[23][24]**. We found that our wasp Gag-Pol protein of interest had homologs in parasitic species, including aphids, mosquitoes, and ticks. Of these species, we were particularly interested in the presence of capsid-like proteins in ticks, given our previous work to develop resources for studying the host-manipulating properties of tick saliva **[25]**. Given that the salivary glands of ticks secrete venom protein homologs into their host upon biting, ticks can be considered venomous species **[26]**. For these reasons, we prioritized one of the tick Gag protein homologs for further analysis as part of our investigation into capsid-like proteins in venoms (XP_042148722.1 from *Ixodes scapularis*) (Table 1).

Evidence of capsid genes horizontally transferred from viruses to *Ixodes* tick species

Because the evolutionary acquisition of viral genes such as capsids is an example of inter-kingdom horizontal gene transfer (HGT), we next deployed a software pipeline specifically tailored to detect inter-kingdom HGT events — preHGT **[10]**. Of the original venomous species included in our list, only nine had gene models available through NCBI (a prerequisite for preHGT) at the time we ran the pipeline. Although the original venomous species dataset we used for HMM-based analyses didn't include tick species, we decided to include the ticks *Ixodes scapularis* and *Ixodes persulcatus* in our HGT analysis because we'd found a capsid protein in ticks in our follow-up BLAST searches, described above.

The preHGT analysis revealed 66 potential horizontal transfer events between our target species and all available genomes on NCBI. Of those, there were 21 instances of putative gene transfer from viruses to our species of interest. 14 of these were to parasitoid wasps, 11 of which were the previously studied bracovirus found within the

Microplitisdemolitor genome. We identified seven viral HGT events outside of wasps, six of which we found in ticks (Table 1). The other non-wasp putative viral HGT event was of a nucleoside transporter in *Sepia pharaonis*, the pharaoh cuttlefish.

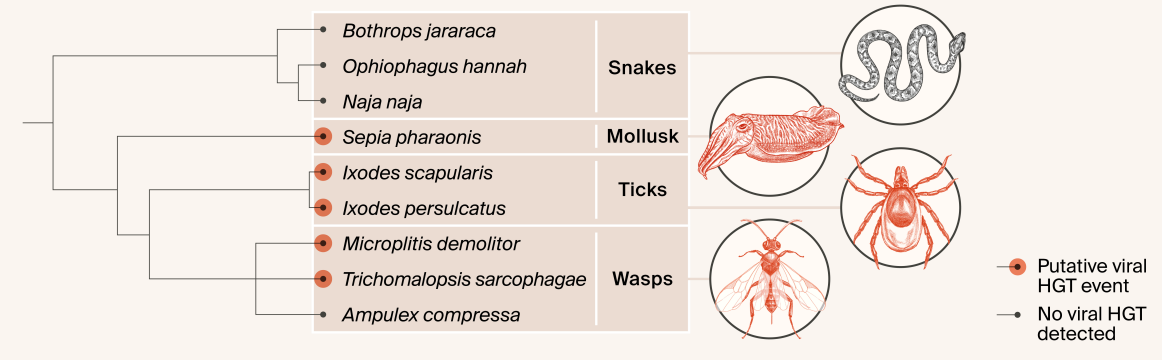

Figure 3

An HGT-based search reveals putative viral HGT events in the genomes of venomous animals.

We used the horizontal gene transfer (HGT) detection tool "preHGT" to identify putative viral HGT events in the genomes of nine venomous animals, including snakes, a cuttlefish, ticks, and parasitoid wasps. Evolutionary relationships between the species we searched are shown on a tree. We identified putative viral HGT events in ticks, parasitoid wasps, and the pharaoh cuttlefish. These events are marked by red circles at the leaf tips. The Newick file we used to generate this visual can be found here, the list of input genomes here, and the hits from this HGT search here.

Four of the six tick viral HGT events were predicted to be capsid-related sequences. These were all annotated as nucleocapsid proteins, or "N proteins," from several different negative-stranded RNA viruses. These RNA viruses belong to groups known to infect ticks, such as Bunyavirales and Rhabdoviridae, yet these types of viruses don't go through a DNA phase in their replication. Integrated RNA viral genes have previously been noted in tick genomes, though their biological significance is unknown **[27]**. It's intriguing to see RNA-derived viral nucleocapsids integrated into DNA-based tick genomes. The two other virus-to-tick HGT events in the dataset were both viral RNA-dependent RNA polymerase (RdRp). Nucleocapsid proteins from these RNA viruses are involved in binding and encapsulating RNA virus genomes and interact with viral RdRp to initiate viral genome replication **[28][29]**. Given these interesting

findings, we decided to prioritize all of these additional tick proteins for further analysis (Table 1).

Protein	Tick species	Closest viral hit	Annotation	Protein family
XP_042148722.1	*Ixodes scapularis*	N/A	Retro-transposon Gag protein	pfam03732: Retro pfam13650:Asp_p pfam00098; zf-C(
KAG0420414.1	*Ixodes persulcatus*	Tahe rhabdovirus	N protein	pfam00945: Rhab
KAG0427517.1	*Ixodes persulcatus*	Tahe rhabdovirus 3	N protein	pfam00945: Rhab
EEC12039.1	*Ixodes scapularis*	Sara tick phlebovirus	N protein	pfam05733: Tenui
EEC17452.1	*Ixodes scapularis*	South Bay sabavirus (bunyavirus order)	N protein	pfam02477: Nairo
KAG0443635.1	*Ixodes persulcatus*	Totiviridae sp.	RdRp	N/A
KAG0444350.1	*Ixodes persulcatus*	Phenuiviridae	RdRp	pfam12603: Bunya like

Table 1

Tick proteins of potential viral origin.

This table lists the seven tick proteins of potential viral origin that we found in this study. We identified the first protein on this list (XP_042148722.1) as a homolog of a viral capsid endogenized by a parasitoid wasp. We identified the other six proteins using HGT-based analysis of animal genomes. This table shows the protein accession number and associated metadata: the tick species in which we found the protein, the virus that the gene probably originated from, and the protein's annotation and protein family. We're also showing the results of our manual genomic context analysis of each gene, where we looked for signs of domestication and salivary expression. We assessed whether each gene is inserted into the tick genome, if it has introns, and if it's expressed in the salivary gland. Last, we highlight the proteins for which we did follow-up analysis using Foldseek and note the associated figure numbers. Annotated GenBank files of the contigs for each protein are available here.

Capsid genes identified in ticks contain introns and are expressed in salivary glands

The HMM- and HGT-based searches led us to identify two different types of potential viral capsids in tick genomes — retrotransposon Gag-Pol capsid-like proteins and negative-stranded RNA viral nucleocapsids and associated RdRps. We next examined the genomic context of each of these genes to look for further evidence of endogenization. All seven genes are located within larger tick genomic contigs, suggesting integration into the tick genome rather than viral contamination that was sequenced along with the tick genome. Six out of seven contained introns (Figure 4), which suggests that the genes have been present in the tick genomes long enough to acquire eukaryotic genome characteristics. This result adds confidence that these sequences aren't derived from viral contaminants in the original datasets but, rather, are endogenized viral genes.

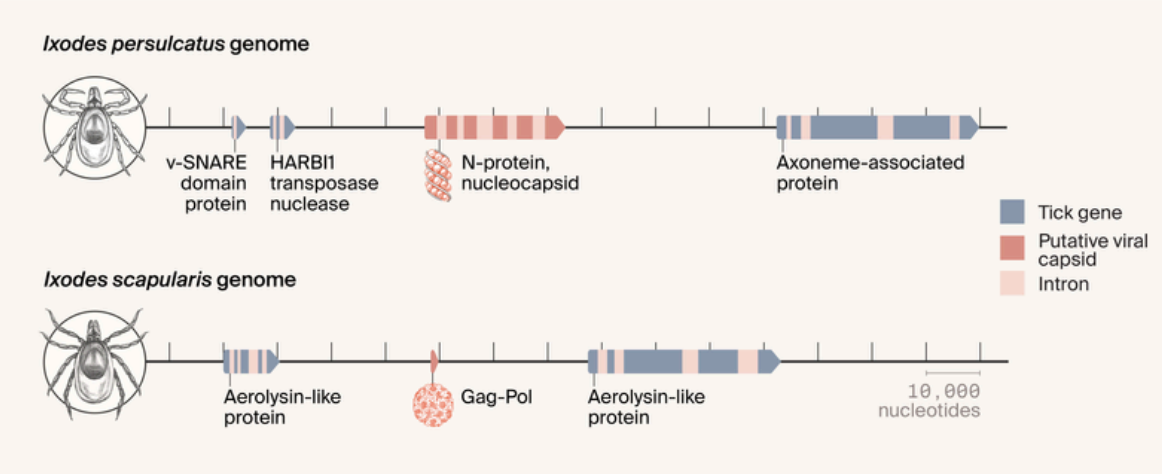

Figure 4

Genomic context of two viral capsid genes endogenized in tick genomes.

Genomic context of the genes encoding KAG0427517.1, the rhabdovirus-derived N protein in *Ixodes persulcatus*, and XP_042148722.1, the Gag-Pol protein in *Ixodes scapularis*. Annotations (determined by looking for related proteins using BLAST, HHpred, and Foldseek) are included below each coding region. Genes of tick origin are grey-blue and viral-derived genes are dark orange. Introns are pale orange bars. Full annotated GenBank files of these contigs can be found here and here.

Because we found most of the viral capsid candidates using the HGT approach rather than the transcriptome analysis, we wondered whether these genes are expressed. We were particularly curious whether they're expressed in tick saliva, as this would suggest that the genes are involved in venom-related functions. We found all of our candidate tick capsids in *Ixodes scapularis* and *Ixodes persulcatus*; however, there's no publicly available *I. persulcatus* salivary transcriptome. There are salivary transcriptomes from *Ixodes scapularis* and *Ixodes ricinus,* so we used those to BLAST the capsid candidates. We found hits to five of the seven candidates, all in the *I. ricinus* salivary transcriptome, suggesting these genes have homologs in *I. ricinus* that are actively produced in the salivary glands.

Searching venoms for proteins with structures similar to viral capsids identifies capsid-like proteins in multiple blood-sucking species and other invertebrates

Due to the rapid sequence evolution of viral genomes, sequence-based approaches can often fail to detect similarity between distantly evolutionarily related genes. However, protein structure can remain conserved even when sequences diverge dramatically **[30]**. For this reason, we next turned to structure-based homology searches to explore whether viral capsids may be present in additional species.

To search for structural homologs of capsid proteins in venoms, we first used Foldseek, a tool that enables structure-based searches for similar proteins **[14]**. We performed structure-based searches on three representative capsid structural proteins (not the RdRps) identified above — the *Ixodes* rhabdovirus nucleocapsid (KAG0427517), the *Ixodes* sabavirus nucleocapsid (EEC17452), and the *Ixodes* Gag-Pol protein (XP_042148722) — as well as the bracovirus major capsid protein from the parasitoid wasp *Microplitis demolitor* (vp39, NP_001401748).

Dealing with disorder in structure comparison

One thing to note is that these proteins all contain disordered regions, which complicates structural interpretation. We get an idea of the amount of disorder in a protein using the pLDDT (predicted local distance difference test), which is calculated by AlphaFold and gives an estimate of the amount of confidence for the structure on a per-residue basis **[31]**. To get an idea of the total amount of disorder for the protein, we average the per-residue pLDDT scores. Regions of proteins that are well-ordered will have a score between 90–100, regions with a score between 70–90 have some structure but have less confidence, regions between 50–70 are generally low-confidence, and regions with a pLDDT of less than 50 are likely disordered **[32]**. KAG0427517 had an average pLDDT of 62; EEC17452 had an average pLDDT of 77; NP_001401748 had an average pLDDT of 72; and XP_042148722 had an average pLDDT of 78. Despite the disorder throughout these proteins, we did get hits that align well with the structured regions of the proteins.

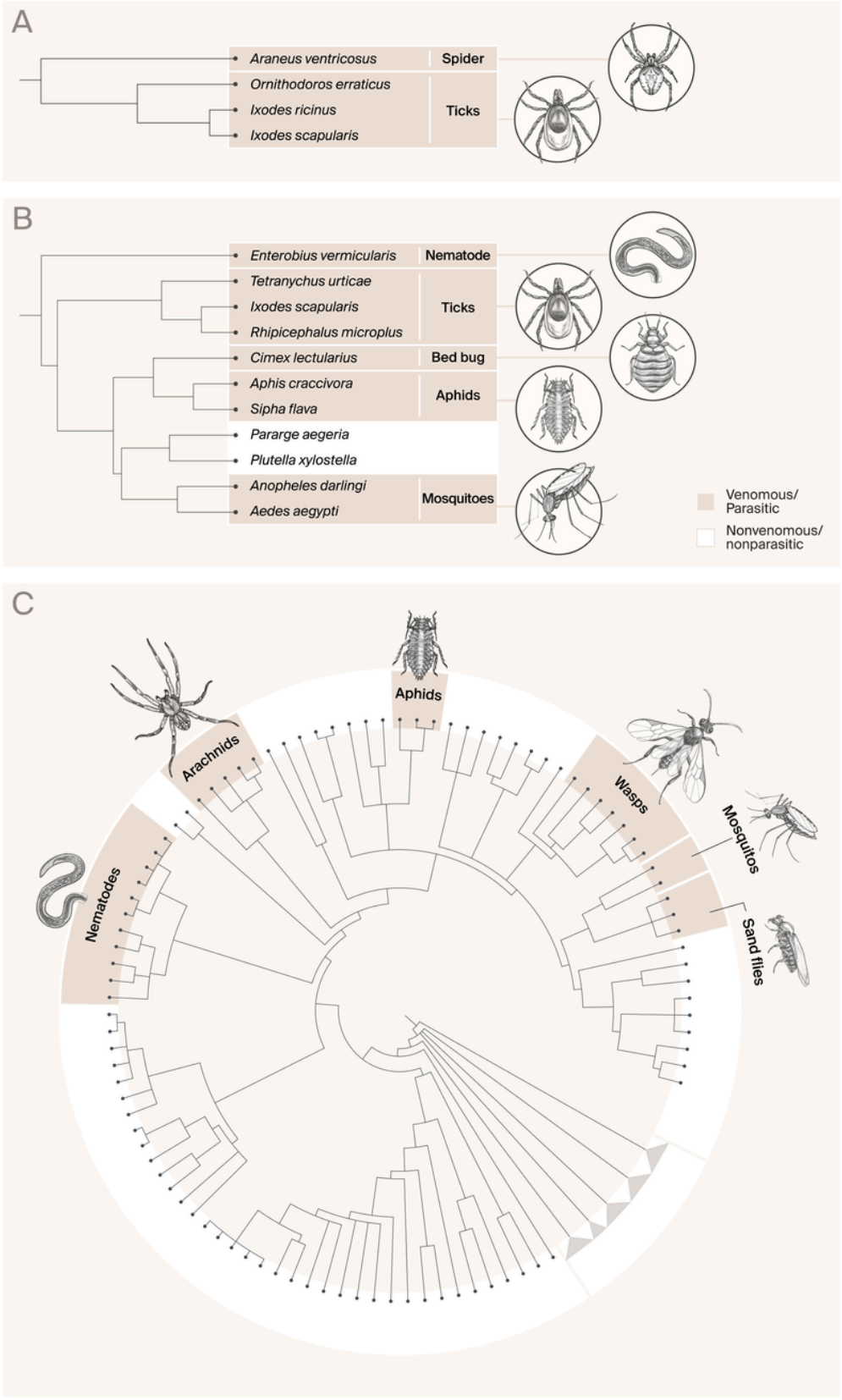

Figure 5

Distribution of structural homologs of endogenized viral proteins from ticks across the AlphaFold database.

We used Foldseek to identify structural homologs of:

(A) the tick sabavirus N protein (EEC17452.1)

(B) the tick rhabdovirus N protein (KAG0427517.1)

(C) the tick Gag-Pol protein (XP_042148722.1). In this panel, we've collapsed all non-animal clades.

We show evolutionary relationships between the species with Foldseek hits to each protein using phylogenetic trees. In this case, every species on the tree has at least one hit to the query tick protein. Clades of animals with parasitic or venomous lifestyles are indicated with brown highlights around leaf tips and name, whereas non-venomous and non-parasitic organisms are shown with white highlighting.

Newick files for trees shown in this figure can be found here. In panel C, we couldn't show individual species names due to the size and complexity of the tree. However, information about all hits for each query, including their species of origin, can be found here.

Sabavirus N hits (EEC17452.1)

The *I. scapularis* nucleocapsid protein EEC17452.1 showed homology to six tick proteins and one spider protein (Figure 5, A). We found the original EEC17452.1 protein plus one additional *I. scapularis* hit. We also had hits in *I. ricinus*; the soft tick, *Ornithodros erraticus*; and the orb-weaver spider, *Araneus ventricosus.*

Rhabdovirus N hits (KAG0427517.1)

We found structural homologs of the *Ixodes persulcatus* rhabdovirus nucleocapsid protein KAG0427517 protein in the ticks *Ixodes scapularis* and *Rhipicephalus microplus*, as well as several other species with intriguing parallels to ticks (Figure 5, B). Namely, we found structural homologs in other hematophagous (blood-sucking) species — mosquitos (*Aedes aegypti* and *Anopheles darlingi*) and bed bugs (*Cimex lectularius*). Intriguingly, we also found homologs in two sap-sucking species — the

aphids *Aphis craccivora* and *Sipha flava*. We also found homologs in the plant parasitic spider mite, *Tetranychus urticae*, and the human parasitic helminth worm, *Enterobius vermicularis*, also known as pinworm. This pattern is compelling because all of these species, including ticks, have parasitic interactions with their host species, suggesting that the viral capsid may play a role in diverse types of parasitic lifestyles. The only non-parasites that had structural homologs of these proteins were the diamondback moth, *Plutella xylostella,* and the speckled wood butterfly, *Pararge aegeria*. Interestingly, *P. xylostella* is a natural host of several parasitoid wasps, suggesting that the nucleocapsid protein could be the result of parasite-to-host HGT. The significance of the N protein homolog in *P. aegeria* is unknown.

Gag-Pol hits (XP_042148722.1)

The tick retrotransposon Gag-Pol protein XP_042148722.1 returned 924 hits in our Foldseek search. These hits are widely distributed across eukaryotes, including many parasitic blood and sap-feeding species (Figure 5, C). This Foldseek search also identified Arc1 (3.69×10^{-8}) and Arc2 (2.00×10^{-5}) from *Drosophila melanogaster* as structural homologs of the tick gal-pol protein. Drosophila Arc proteins are domesticated viral capsids that participate in RNA delivery between cells **[23]**.

Bracovirus vp39 hits (NP_001401748)

The only structural homologs we found of the bracovirus vp39 protein NP_001401748 are from other parasitoid wasps, *Cotesia chilonis* and *Chelonus inanitus*.

Co-clustering venom proteins with known viral capsids reveals a putative, uncharacterized eukaryotic viral capsid

We next wanted to see if we could extend structural homology-based approaches to whole-venom transcriptomes. Because all of our prior results highlighted the presence of potential capsids in tick species, we chose to focus on proteins from tick salivary transcriptomes for this analysis. We analyzed structures from *Ornithodoros turicata*, a species of soft-bodied tick in the family Argasidae. We aggregated folded protein structures from *O. turicata* and folded protein structures from known capsid proteins

from the VOG database. We then performed clustering to see if any capsids would cluster with proteins from tick saliva. We identified 47 structural clusters (Figure 6, A). From this starting point, we could explore the candidate list of clusters for possible evidence of viral proteins clustering with venom proteins. We see the majority of clusters are quite specific either to viruses or to *O. turicata*, but there were six candidate clusters containing both a mix of proteins from the two datasets and a high degree of structural homology (TM > 0.5).

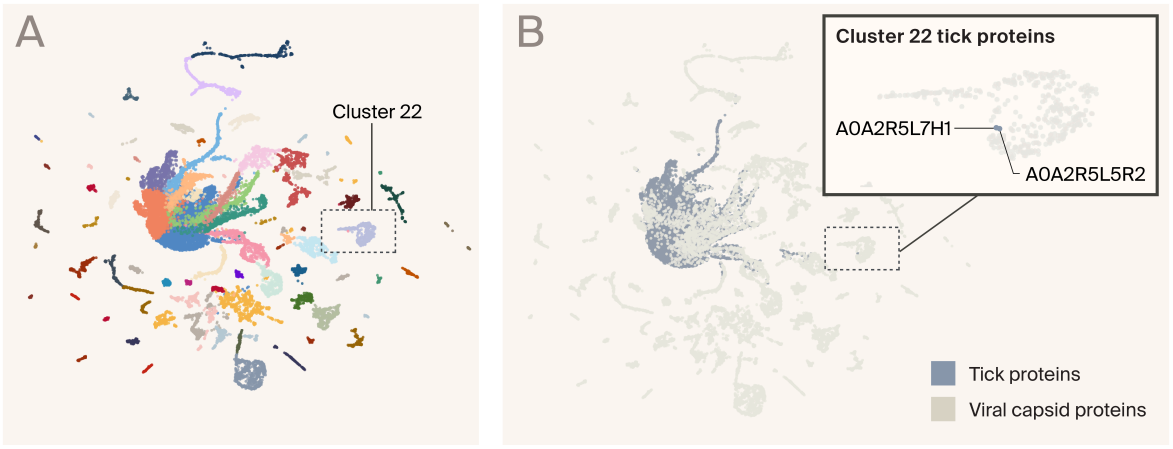

Figure 6

ProteinCartography analysis of viral capsid proteins with *Ornithodoros turicata* proteins reveals two tick proteins that cluster with capsid proteins.

We clustered all *O. turicata* structures available on UniProt with ESMFold predictions of the capsid proteins in the VOG database using ProteinCartography.

(A) A UMAP of co-clustered tick and viral proteins. Here, each protein is a single point on the UMAP and is colored by cluster membership.

(B) The same UMAP as in (A), but here tick proteins are green, and viral proteins are grey. Generally, the capsid proteins and tick proteins are clustered separately). We identified cluster 22 (boxed) as a cluster that contained two tick proteins alongside proteins from four different viral capsid VOGs.

(C) Zoom-in on cluster 22. Tick proteins A0A2R5L5R2 and A0A2R5L7H1 cluster with VOG proteins, some of which are annotated as mimivirus capsids.

We followed up on cluster 22, which contained tick proteins A0A2R5L5R2 and A0A2R5L7H1, along with proteins annotated as “major capsid protein” from mimivirus (Figure 6, B). Mimiviruses are giant viruses that typically infect species of amoeba **[33]**. However, the mimivirus capsid protein is in the same protein family as the capsid from African swine fever virus, which *O. turicata* ticks transmit. By sequence, the *O. turicata* proteins found in this cluster are highly divergent from capsids from both the mimivirus and African swine fever virus. Unfortunately, there are no *O. turicata* genomes available, so it’s not currently possible to see if this protein comes from an unknown contaminating virus or is endogenous to the tick genome.

After our initial test with *O. turicata*, we didn't pursue this clustering approach further. This method is computationally challenging to scale outside of a few salivary transcriptomes, and our ability to interpret the hits that come out of this approach is limited.

Key takeaways

- We detected putative endogenized viral capsids encoded by ticks and several other parasitic species, and the tick-encoded viral capsids are expressed in tick saliva.
- Our highest-confidence tick-encoded capsid types are RNA nucleocapsids and retrotransposon Gag-Pol proteins. We predict that both interact with and help package RNA molecules.
- These results suggest the possibility that some parasites use packaged nucleic acids to interact with their hosts.

Discussion

Venomous species have evolved an enormous biological arsenal of host-manipulating factors. This particular study was inspired by the ancient use of viral capsids for host gene delivery by parasitoid wasps. We took a multi-pronged search across venomous species, looking for examples where other venomous species may have evolved the use of viral capsids. The two major categories of viral capsids we identified originate from RNA viruses [rhabdovirus and sabavirus (a type of bunyavirus)] and the retrotransposon Gag-Pol protein from LTR retrotransposons.

We were excited to detect these capsid-like proteins outside of parasitoid wasps (Figure 1). Taxonomic analysis of these species suggests that some endogenized capsids are enriched in organisms that have parasitic interactions with their hosts. In particular, several tick species appear to have homologs of multiple types of viral capsids. These appear to have been endogenized in the ticks' genomes and expressed in their saliva, the tick equivalent of venom. Integration of viral sequences into tick genomes has been observed before **[34][35][27]**. Hard ticks feed on hosts for days to weeks, unlike the rapid predator-prey interactions found in many venomous species, such as snakes. This extended host-parasite interaction could be analogous to the extended interaction of parasitoid wasps and lepidopteran hosts. This sustained interaction likely requires modes of host manipulation that act over longer timescales — gene delivery may be one of these.

Our findings suggest that the endogenization of viral capsids by venomous and/or parasitic species may be more widespread than previously understood, and these endogenized capsids are common in parasitic species that have extended interactions with their hosts. Right now, we don't know the physiological role, if any, of these endogenized capsids in the parasitic lifestyle. However, we believe our findings are an intriguing starting point for further study into the specific roles of these capsids in host–parasite biology and may suggest novel approaches to molecular delivery in mammals.

Next steps

Our initial work has established that capsid-like proteins are found in a variety of parasitic species, including some that bite humans. The biological significance of these endogenized capsids is still unknown. For us, the most translationally exciting outcome would be that capsids from human parasites are participating in gene delivery to human cells. In our own follow-up work, we took a stab at identifying potentially packaged DNA molecules in parasites by looking for signatures of circularized DNA in whole-genome sequencing libraries **[3]**. We were unable to devise a way to use RNA sequencing libraries to identify potentially packaged RNAs, which would have more directly built on our finding here. The results from our circularized DNA analysis were interesting, but not sufficiently compelling to warrant further investment from us.

References

1. Dupuy C, Huguet E, Drezen J-M. (2006). Unfolding the evolutionary story of polydnaviruses. https://doi.org/10.1016/j.virusres.2006.01.001

2. Glatz RV, Asgari S, Schmidt O. (2004). Evolution of polydnaviruses as insect immune suppressors. https://doi.org/10.1016/j.tim.2004.10.004

3. Borges AL, Celebi FM, Cooper RO, McDaniel EA. (2024). Identifying circular DNA using short-read mapping. https://doi.org/10.57844/ARCADIA-AD7F-7A6D

4. Wang D, Tai PWL, Gao G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. https://doi.org/10.1038/s41573-019-0012-9

5. Gauthier J, Boulain H, van Vugt JJFA, Baudry L, Persyn E, Aury J-M, Noel B, Bretaudeau A, Legeai F, Warris S, Chebbi MA, Dubreuil G, Duvic B, Kremer N, Gayral P, Musset K, Josse T, Bigot D, Bressac C, Moreau S, Periquet G, Harry M, Montagné N, Boulogne I, Sabeti-Azad M, Maïbèche M, Chertemps T, Hilliou F, Siaussat D, Amselem J, Luyten I, Capdevielle-Dulac C, Labadie K, Merlin BL, Barbe V, de Boer JG, Marbouty M, Cônsoli FL, Dupas S, Hua-Van A, Le Goff G, Bézier A, Jacquin-Joly E, Whitfield JB, Vet LEM, Smid HM, Kaiser L, Koszul R, Huguet E, Herniou EA, Drezen J-M. (2021). Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization. https://doi.org/10.1038/s42003-020-01623-8

6. Drezen J-M, Leobold M, Bézier A, Huguet E, Volkoff A-N, Herniou EA. (2017). Endogenous viruses of parasitic wasps: variations on a common theme. https://doi.org/10.1016/j.coviro.2017.07.002

7. Wu X, Wu Z, Ye X, Pang L, Sheng Y, Wang Zehua, Zhou Y, Zhu J, Hu R, Zhou S, Chen J, Wang Zhizhi, Shi M, Huang J, Chen X. (2022). The Dual Functions of a Bracovirus C-Type Lectin in Caterpillar Immune Response Manipulation. https://doi.org/10.3389/fimmu.2022.877027

8. Herniou EA, Huguet E, Thézé J, Bézier A, Periquet G, Drezen J-M. (2013). When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. https://doi.org/10.1098/rstb.2013.0051

9. Finn RD, Clements J, Eddy SR. (2011). HMMER web server: interactive sequence similarity searching. https://doi.org/10.1093/nar/gkr367

10 Dutton RJ, Reiter T. (2024). PreHGT: A scalable workflow that screens for horizontal gene transfer within and between kingdoms. https://doi.org/10.57844/ARCADIA-JFBP-7P11

11 Reiter T. (2024). Clustering the NCBI nr database to reduce database size and enable faster BLAST searches. https://doi.org/10.57844/ARCADIA-W8XT-PC81

12 Gladyshev EA, Meselson M, Arkhipova IR. (2008). Massive Horizontal Gene Transfer in Bdelloid Rotifers. https://doi.org/10.1126/science.1156407

13 Rancurel C, Legrand L, Danchin E. (2017). Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. https://doi.org/10.3390/genes8100248

14 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1038/s41587-023-01773-0

15 Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. (2022). ColabFold: making protein folding accessible to all. https://doi.org/10.1038/s41592-022-01488-1

16 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

17 Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. (2023). Clustering predicted structures at the scale of the known protein universe. https://doi.org/10.1038/s41586-023-06510-w

18 Avasthi P, Bigge BM, Celebi FM, Cheveralls K, Gehring J, McGeever E, Mishne G, Radkov A, Sun DA. (2024). ProteinCartography: Comparing proteins with structure-based maps for interactive exploration. https://doi.org/10.57844/ARCADIA-A5A6-1068

19 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574

20 Letunic I, Bork P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. https://doi.org/10.1093/nar/gkab301

21 Dutton RJ, McDaniel EA, Morin M. (2024). Identifying candidate accessory domains by mining putative venom protein fusions. https://doi.org/10.57844/ARCADIA-D3DC-7499

22 Havecker ER, Gao X, Voytas DF. (2004). https://doi.org/10.1186/gb-2004-5-6-225

23 Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. https://doi.org/10.1016/j.cell.2017.12.024

24 Segel M, Lash B, Song J, Ladha A, Liu CC, Jin X, Mekhedov SL, Macrae RK, Koonin EV, Zhang F. (2021). Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. https://doi.org/10.1126/science.abg6155

25 Chou S, Poskanzer KE, Thuy-Boun PS. (2024). Robust long-read saliva transcriptome and proteome from the lone star tick, Amblyomma americanum. https://doi.org/10.57844/ARCADIA-3HYH-3H83

26 Cabezas-Cruz A, Valdés JJ. (2014). Are ticks venomous animals? https://doi.org/10.1186/1742-9994-11-47

27 Martyn C, Hayes BM, Lauko D, Mithun E, Castañeda G, Bosco-Lauth A, Kistler A, Pollard KS, Chou S. (2022). mNGS Investigation of Single Ixodes pacificus Ticks Reveals Diverse Microbes, Viruses, and a Novel mRNA-like Endogenous Viral Elements. https://doi.org/10.1101/2022.08.17.504163

28 Ding B, Qin Y, Chen M. (2016). Nucleocapsid proteins: roles beyond viral <scp>RNA</scp> packaging. https://doi.org/10.1002/wrna.1326

29 Luo M, Terrell JR, Mcmanus SA. (2020). Nucleocapsid Structure of Negative Strand RNA Virus. https://doi.org/10.3390/v12080835

30 Sykes J, Holland BR, Charleston MA. (2022). A review of visualisations of protein fold networks and their relationship with sequence and function. https://doi.org/10.1111/brv.12905

31 Mariani V, Biasini M, Barbato A, Schwede T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. https://doi.org/10.1093/bioinformatics/btt473

32 Akdel M, Pires DEV, Porta Pardo E, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, Shenoy A, Zhu W, Kundrotas P, Ruiz Serra V, Rodrigues CHM, Dunham AS, Burke D, Borkakoti N, Velankar S, Frost A, Lindorff-Larsen K, Valencia A, Ovchinnikov S, Durairaj J, Ascher DB, Thornton JM, Davey NE, Stein A, Elofsson A, Croll TI, Beltrao P. (2021). A structural biology community assessment of AlphaFold 2 applications. https://doi.org/10.1101/2021.09.26.461876

33 Scola BL, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie J-M, Raoult D. (2003). A Giant Virus in Amoebae. https://doi.org/10.1126/science.1081867

34 Bell-Sakyi L, Attoui H. (2013). Endogenous tick viruses and modulation of tick-borne pathogen growth. https://doi.org/10.3389/fcimb.2013.00025

35 Russo AG, Kelly AG, Enosi Tuipulotu D, Tanaka MM, White PA. (2019). Novel insights into endogenous RNA viral elements in Ixodes scapularis and other arbovirus vector genomes. https://doi.org/10.1093/ve/vez010