Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A tiered metatranscriptomic survey of viral diversity associated with Leptoglossus phyllopus and its gut symbionts

Patrick T. Stillson

pstillson@fsmail.bradley.edu

Bradley University

Research Article

Keywords: Caballeronia, Heteroptera, Leptoglossus phyllopus, Symbiosis, Virome

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11089388/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
No competing interests reported.

Table 1 is available in the Supplementary Files section.

Title: A tiered metatranscriptomic survey of viral diversity associated with *Leptoglossus phyllopus* and its gut symbionts

Patrick T. Stillson[a,1,*]

[a] Department of Biology, Emory University, Atlanta, GA, USA

[1] Current Address: Biology Department, Bradley University, Peoria, IL, USA

* Corresponding Author: Patrick T. Stillson, pstillson@fsmail.bradley.edu

ORCID:

Patrick T. Stillson: 0000-0001-5382-7704

Key Words:

Caballeronia, Heteroptera, *Leptoglossus phyllopus*, Symbiosis, Virome

Abstract

Viruses are increasingly recognized as important components of insect-associated microbial communities, yet the viromes of many non-model insects remain poorly characterized. Here, I conducted an exploratory survey of viral diversity associated with the eastern leaf-footed bug, *Leptoglossus phyllopus*, and two strains of its environmentally acquired *Caballeronia* gut symbiont. Using transcriptomic libraries together with host and symbiont genome assemblies, I applied a tiered viral discovery pipeline integrating dedicated viral identification tools, protein homology searches, hidden Markov model-based detection of viral-like domains, and genomic screening for endogenous viral elements (EVEs). I identified 20 high-confidence viral contigs representing multiple viral taxonomic groups, including Caudoviricetes, Orthototiviridae, and several plant-associated viral taxa. Broader homology searches recovered more than 200 additional viral candidates, including many with low similarity to characterized viruses, while HMM-based screening identified 2,714 contigs containing viral-like protein domains. Genomic analyses revealed 24,796 candidate EVE loci, the majority of which were associated with repetitive genomic regions. Of 2,468 viral-like transcripts that mapped to the host or symbiont genomes, 1,518 overlapped candidate EVE loci and were classified as transcribed EVEs, while 950 remained stronger candidates for exogenous viral sequences. These results demonstrate that the viral landscape associated with *L. phyllopus* includes a mixture of candidate exogenous viruses, bacteriophages, environmentally acquired viral sequences, and transcriptionally active EVEs. This study establishes an initial reference for the *L. phyllopus* virome and highlights the value of integrating transcriptomic and genomic evidence when characterizing viral diversity in insect-microbe systems.

Introduction

Insects represent the most diverse group of animals and contribute to numerous ecosystem services, including acting as a food source, pollinating crops, and cycling nutrients [1]. Within this group, heteropteran true bugs include many agriculturally important pests that directly impact global crop production [2]. Of these, the Eastern leaf-footed bug (*Leptoglossus phyllopus* (Linnaeus) (Hemiptera: Coreidae)), is a widely distributed polyphagous herbivore capable of feeding on a broad range of economically important crops including legumes, tomatoes, tree nuts, and citrus [3,4]. Despite its broad geographic distribution and pest status, relatively little is known about its non-bacterial microbial community.

Like many heteropterans, *L. phyllopus* harbors specialized bacterial symbionts (*Caballeronia* spp.) in specialized pouches, termed crypts, in the midgut that are essential for nutrient supplementation and development [5,6]. Unlike the vertically transmitted symbionts (i.e. maternally inherited microbes) common in many other insect systems, the bacterial symbionts of Coreidae, and related taxa, are environmentally acquired each generation [5,7–9]. This mode of symbiont acquisition introduces numerous avenues in which foreign microbes and viruses can interact. With hosts probing the environment early in development for their symbionts, they also acquire foreign non-symbiotic microbes, along with *Caballeronia*, which can both introduce bacteria-associated viruses, and hosts could also acquire insect-associated viruses through this process. These dynamic insect-microbial systems may support complex viral communities, including viruses that infect the insect host, viruses associated with bacterial symbionts, and viruses that persist as genomic remnants of past infections within host or symbiont genomes.

Advances in high-throughput sequencing have greatly expanded our understanding of the diversity of insect-specific viruses (ISVs). Metatranscriptomic studies have revealed that insects harbor highly diverse communities of RNA and DNA viruses [10]. These ISVs can have the potential to act as symbionts, to influence host physiology and immunity, and to interact with microbial symbionts [10–13]. Recent studies in heteropteran insects have uncovered diverse viral communities dominated by RNA viruses, particularly those within the families *Dicistroviridae*, *Flaviviridae*, *Iflaviridae*, and *Rhabdoviridae* [14–17]. However, the viromes of many non-model heteropterans remain poorly characterized [17], limiting our broader understanding of insect-associated viral diversity and the potential roles of viruses in shaping host-symbiont interactions.

A challenge in virome studies is distinguishing actively replicating viruses from endogenous viral elements (EVEs). These EVEs are viral sequences that have become integrated into the host's genome through past viral infections and can persist over evolutionary timescales [18]. Because some EVEs remain transcriptionally active, RNA sequencing alone may incorrectly classify these sequences as evidence of active viral infections [17,19]. Integrating transcriptional data with genomic analyses, such as identifying integration sites and associated repetitive elements, is therefore essential to accurately differentiate active viruses from genomically integrated viral relics [20].

Here, I conducted an exploratory survey of viruses associated with *L. phyllopus* and two strains of its environmentally acquired bacterial symbiont, *Caballeronia*, to characterize a continuum of viral signals. Using metatranscriptome assemblies derived from host gut samples along with host and symbiont genome assemblies, I implemented a multi-step bioinformatic pipeline to 1) identify putative viral sequences, 2) quantify viral transcriptional activity, and 3) detect endogenous viral elements within host and symbiont genomes. By integrating viral discovery tools with protein homology searches and genomic integration analyses, I provide an initial characterization of the viruses associated with this host-symbiont system. These results establish a baseline for understanding the viral diversity in *L. phyllopus* and highlight the complexity of viral interactions within insect–symbiont complexes.

Materials and Methods

Viral Transcriptome Assembly. To identify actively transcribed viral sequences associated with lab-reared *L. phyllopus* and two *Caballeronia* strains of its environmentally acquired bacterial symbiont, a total of 18 transcript libraries were analyzed (nine per symbiont strain). Each RNA library consisted of total RNA pooled from 10 host guts, harboring either *Caballeronia* strain LZ003 or LZ019 (strain LZ003 SRA accessions: SRX23992087 – SRX23992095; strain LZ019 SRA accessions: SRX23992141 – SRX23992149).

Reads were trimmed with Cutadapt v.5.2 [21] to remove adapter sequences, and reads with N > 10%, reads with low quality scores, or reads shorter than 20 bp. After cleanup, reads were mapped to a concatenated genome file, consisting of the host and both symbiont genomes (*L. phyllopus* accession: GCA_041002905.1; LZ003: GCF_031451465.1; LZ019: GCF_031450825.1), using HISAT2 v.2.2.2 [22] with default settings. Unaligned reads were kept, and subsequently assembled *de novo* using MEGAHIT v.1.2.9 [23] with a minimum contig length of 200 bp to ensure recovery of small or highly divergent RNA viruses that may have low coverage.

To maximize sequencing depth and sensitivity for rare or divergent viral sequences, unmapped reads from all 18 libraries were combined prior to *de novo* assembly. Consequently, downstream viral discovery was performed on a pooled metatranscriptome rather than on sample-specific assemblies. This assembly served as the basis for identifying candidate viral sequences.

Detection of Viral Sequences from Transcript Data. Putative viral sequences were identified from the assembled viral transcriptome using two viral discovery tools: VirSorter2 v.2.2.4 (reference database v.0.4) [24] and VIBRANT v.1.2.1 (reference database v.1.2.1) [25]. These programs identify viral sequences based on gene content, viral hallmark genes, and machine-learning classification of viral genomic signatures. To assess completeness and remove potential false positives that may have been produced by each tool, the candidate viral sequences were subsequently evaluated using CheckV v.1.0.3 [26]. Only sequences classified as medium quality or higher were retained, and viral contigs sharing > 95% sequence identity were collapsed to produce a non-redundant set of 20 viral contigs (Tier 1 of Fig. 1).

To assign putative viral taxonomy, the curated contigs were queried against the NCBI non-redundant protein database (NCBI NR 2023-07-28) using DIAMOND (BLASTx) v.2.1.22 [27]. DIAMOND results were filtered to only retain hits taxonomically assigned to "Viruses." For each of the remaining contigs, the top-scoring viral hits were retained for downstream annotation.

To estimate transcriptional activity of the 20 identified viral contigs, reads were mapped back to the curated viral contig set using BBMap v.39.08 [28] with minimum alignment identity set to 0.90. Alignment files were processed using bedtools Genome Coverage v.2.31.1 [29] to calculate read depth, coverage, and raw read counts for each viral contig. To enable comparisons of viral expression levels across contigs, read counts were normalized to transcripts per million (TPM), providing a relative measure of viral transcript abundance within the metatranscriptome dataset.

Viral Protein Homology Search: Similarity-Based Identification. Using the assembled viral transcriptome, contigs were compared to the NCBI NR Database (2023-07-28) using DIAMOND (BLASTx). For each contig, only the top-scoring hit, that was taxonomically assigned to "Viruses," was retained after filtering based on the following thresholds: E-value $\leq 1e^{-5}$, bit score ≥ 50, alignment length ≥ 80 aa, and percent identity $\geq 25\%$. This permissive identity threshold was selected to enable detection of divergent or potentially novel viral sequences, with higher identity values interpreted as evidence of closer similarity to well-characterized viruses. This provided 207 putative viral contigs (Tier 2 of Fig. 1).

Viral Protein Homology Search: HMM-Based dentification. Contigs from the assembled viral transcriptome that lacked viral taxonomic assignment from DIAMOND were subsequently translated into amino acid sequences using Transeq v.5.0.0 [30], to retain predicted open reading frames (ORFs). Translated sequences shorter than 50 amino acids or longer than 100,000 amino acids were removed to exclude spurious ORFs and assembly artifacts. Translated protein sequences were screened for conserved viral-like protein domains using hmmsearch v.3.4 [31] against a Reference Viral Database (RVDB) v.30.0 [32] using an E-value $\leq 1e^{-10}$ to identify significant matches. The HMMER output was subsequently filtered to only retain contigs meeting a significance threshold of E-value $\leq 1e^{-20}$,

domain score ≥ 100, and domain coverage ≥ 50%. For each contig, only the highest domain score was retained. This provided 2,714 contigs containing viral-like protein domains (Tier 3 of Fig. 1).

Identification and Characterization of Endogenous Viral Elements. To identify whether the previously identified 2,714 contigs containing viral-like protein domains may be associated with viral integration, I investigated the presence of EVEs that may be integrated within the host or symbiont genomes. The concatenated genome assemblies were translated into amino acid sequences using ORFipy v.0.0.4 [33] across both strands and predicted ORFs shorter than 100 amino acids or longer than 10,000 amino acids were removed to exclude spurious ORFs and assembly artifacts. Resulting protein sequences were screened for conserved viral protein domains using hmmsearch against the RVDB. The output was filtered to retain only domains with an i-Evalue ≤ $1e^{-10}$, domain score ≥ 50, and domain coverage ≥ 30%. For each contig, only the highest-scoring viral domain hit was retained. Candidate viral ORFs identified by HMMER were mapped back to the concatenated genome assemblies using NCBI BLAST+ tblastn v.2.16.0 [34] to determine genomic coordinates of the putative viral integrations. Resulting alignments were filtered to retain matches with E-value ≤ $1e^{-10}$, bit score ≥ 80, and alignment length ≥ 80 amino acids. For each ORF, only the top-scoring genomic match was retained. To account for adjacent integration-associated features, flanking regions extending 1,000 bp upstream and downstream of each locus were included.

To evaluate whether candidate EVEs were associated with repetitive genomic features commonly linked to viral integration, a custom repeat library was first generated from the concatenated genome assemblies using RepeatModeler v.2.0.5 [35]. The genomes were subsequently screened against this custom repeat library using RepeatMasker v.4.1.5 [36] to identify transposable elements and repetitive regions. Repeat annotations were intersected with EVE loci, and any EVE loci overlapping with one or more repeat annotations were classified as repeat associated. These repeat-associated loci were interpreted as genomic regions potentially enriched for transposable elements or other repetitive features associated with viral integration events.

To characterize the transcriptional activity of putative EVEs, the 2,714 transcript-derived viral contigs identified above under the "Viral Protein Homology Search" pipeline were mapped back to the concatenated genome assemblies using NCBI BLAST+ blastn v.2.16.0 [34]. Overlap between transcript-derived viral coordinates and repeat-associated EVE loci was assessed. Viral sequences were categorized as "likely exogenous viruses" if there was evidence in the transcript assemblies but there was no overlap with the EVE loci, or "transcribed EVEs" if the EVE loci overlapped with the transcript region.

Results

High-Confidence Viral Sequences Recovered from Transcriptome Data. I identified 20 non-redundant viral contigs that met stringent quality criteria following evaluation with CheckV. These contigs ranged from 2,578 – 40,577 bp in length and were assigned across multiple viral taxonomic groups based on DIAMOND similarity searches against the NCBI NR database. Viral sequences were most frequently associated with Caudoviricetes (6 contigs) and Orthototiviridae (4), although several contigs exhibited low similarity to previously characterized viruses, suggesting the presence of divergent viral lineages (Fig 2A-B, Table 1, Table S1).

Read mapping revealed substantial variation in transcript abundance among viral contigs. The TPM values ranged from 6,872.87 to 1,285,805, indicating that all recovered viral contigs were highly represented within the transcriptome dataset (Fig 2C). The most abundant viral contigs were associated with Picornavirales and Totiviridae, whereas other viral sequences were detected at comparatively low abundance. Together, these results indicate the presence of multiple viral taxa represented by abundant transcripts within *L. phyllopus* gut-associated communities.

Homology-Based Searches Reveal Additional Viral Diversity. To expand viral discovery beyond high-confidence viral genome predictions, assembled transcript contigs were screened for protein similarity against the NCBI NR database. After filtering, 207 contigs retained top-scoring matches to viral proteins, representing a >10-fold increase relative to the 20 high-confidence viral contigs identified through genome-based prediction approaches. These viral contigs were associated with a broad range of viral taxa, including Nimaviridae (38 contigs), Parvoviridae (27), Orthototiviridae (18), and Eupolintoviridae (14). Sequence similarity varied considerably among contigs, with some exhibiting high amino acid identity to previously characterized viruses (58 contigs > 80% identity to known viral proteins) and others displaying only distant similarity (63 contigs < 40% similarity), consistent with potentially divergent viral lineages (Fig 3A-D, Fig S1, Table S2).

Viral-Like Protein Domains Identify Additional Candidate Viral Genetic Elements. Because highly divergent viruses may lack detectable sequence similarity to known viral genomes, transcript-derived contigs lacking viral taxonomic assignments were screened for conserved viral protein domains using an HMM-based approach. This analysis identified 2,714 contigs containing significant matches to viral-like protein domains.

Of all annotated protein products identified from the candidate contigs, approximately 10% corresponded to functions commonly found in viral genomes or mobile genetic elements, including reverse transcriptases, polymerases, and helicases (Fig S2, Table S3). However, because these functions are not virus-specific and also occur in eukaryotic and bacterial genomes, as well as transposable elements, the biological origin of these contigs remains unresolved. Their prevalence nevertheless suggests that a substantial fraction of the assembled transcriptome contains sequences with similarity to genes commonly associated with viruses and other mobile genetic elements, greatly exceeding the number of sequences that could be confidently assigned to known viral taxa.

Endogenous Viral Elements are Abundant and Strongly Associated with Repetitive Regions. To determine whether viral-like sequences detected in transcriptomic data corresponded to EVEs, host and symbiont genome assemblies were screened for EVE loci. I identified 24,796 candidate EVE loci distributed across the three genomes (23,975 from *L. phyllopus*, 467 from LZ003, and 354 from LZ019). The majority of candidate EVEs (24,235 loci; 97.74%) overlapped with repetitive genomic regions, while only 561 loci occurred outside annotated repeats, demonstrating a strong association between EVE loci and repetitive DNA (Fig 4A, Fig S3, Table S4).

To distinguish between actively transcribed viruses and endogenous viral elements identified in our list of the 2,714 viral-like domains, transcript-derived viral contigs were compared against the list of 24,796 candidate EVE loci. A total of 2,468 contigs (90.9%) could be confidently mapped to the *L. phyllopus* and symbiont genomes, while 246 transcripts could not be reliably aligned and were excluded from downstream analysis (Fig 4B). Of these mapped transcripts, 1,518 overlapped with the EVE loci and were therefore classified as transcribed EVEs, whereas

950 did not overlap and were classified as likely exogenous viral transcripts and may represent active or recent viral infections rather than genomic integrations (Fig 4C, Table S5).

These sequences indicate that a subset of integrated viral elements remains transcriptionally active and could potentially be misidentified as active viral infections if transcriptomic evidence alone were considered. Taken together, these analyses partitioned viral-associated sequences into exogenous viral candidates, and transcriptionally active EVEs, revealing multiple sources of viral signal within the *L. phyllopus* host-symbiont system.

Discussion

Here, I provide an exploratory characterization of the viral diversity associated with *L. phyllopus* and two strains of its *Caballeronia* symbiont. I detected 20 high-confidence viral contigs using dedicated viral discovery and quality assessment tools, and detected more than 200 additional viral candidates using the broader protein homology searches. The HMM-based search, in addition, identified nearly 3,000 contigs containing viral-like protein domains. The increase in candidate sequences across these levels is consistent with the incompletely characterized diversity of viruses associated with insects, particularly viruses that may be highly divergent from available reference genomes [17,37]. However, through our genomic analyses, I demonstrated that a substantial proportion of this broader viral-like signal may be attributed to transcriptionally active EVEs. Therefore, the tiers used here are best interpreted as a progression from higher-confidence viral sequences (Tier 1) toward increasingly exploratory viral-associated genetic elements (Tier 3), rather than as an estimate of the absolute number of viruses associated with *L. phyllopus* and *Caballeronia*. Collectively, these results suggest that the diversity of viral sequences associated with *L. phyllopus* extends beyond what can be identified through dedicated viral discovery tools, while also highlighting the importance of distinguishing exogenous viruses from other sources of viral-like sequence such as EVEs.

The broad homology searches illustrate both the value and limitations of sequence-based viral discovery. Many Tier 2 contigs exhibited strong similarity to previously characterized viral proteins, while 63 contigs had less than 40% amino acid identity to their best database match. These low-identity sequences may represent highly divergent viruses, which would be expected given the substantial amount of uncharacterized viral diversity present in arthropods [37]. At the same time, decreasing sequence similarity increases uncertainty in both taxonomic assignment and biological origin. This issue was particularly apparent in Tier 3, where many of the identified protein domains represented reverse transcriptases, polymerases, helicases, and other functions that occur in viruses but are not necessarily virus specific. These proteins can also occur in transposable elements, and in both bacterial and eukaryotic genomes. For this reason, the Tier 3 results provide evidence for a large pool of viral-like and mobile genetic elements rather than evidence for thousands of additional viruses. The difference among the three tiers demonstrates the tradeoff in exploratory virome studies: increasing sensitivity improves the ability to recover highly divergent viral sequences, but this broader detection comes with reduced specificity and confidence in detecting viral origin.

Comparison with the host and symbiont genomes helped resolve some of this uncertainty by identifying viral-like transcripts associated with candidate EVE loci. Although I do not consider these EVE-associated sequences evidence of exogenous viruses, their transcription is itself notable because EVEs are widespread in

arthropods and, in some systems, have been associated with antiviral and other host biological functions [38]. Whether the EVEs identified here have functional roles in *L. phyllopus* remains unknown, but their abundance is indicative of them being components of the virome in this system. In contrast, the 950 mapped viral-like transcripts that did not overlap with the candidate EVE loci represent stronger candidates for exogenous viral sequences. However, the absence of genomic overlap alone does not demonstrate active viral replication, and these sequences should therefore remain conservatively interpreted as candidate exogenous viruses.

The exploratory design of this study places several limits on the conclusions that can be drawn from the recovered viral sequences. Because reads from all 18 transcriptomic libraries were combined prior to assembly, the analyses characterize the overall virome associated with the *L. phyllopus* gut, but I cannot determine viral prevalence or variation among individual biological samples or between *Caballeronia* strains. In addition, sequence-based associations cannot establish the biological origin of individual viruses. The recovered sequences may originate from *L. phyllopus*, its microbiota, or ingested plant material, and so more targeted approaches, such as viral isolation or experimental infection, would be required to confirm their biological host and replication. Nevertheless, the combined genomic, transcriptomic, and homology-based analyses provide multiple lines of evidence for a diverse collection of viral sequences associated with this system.

Host-associated Viral Elements. Integrating genomic and transcriptomic evidence allowed us to better distinguish exogenous viral candidates from transcriptionally active endogenous viral elements. RNA sequencing is useful for detecting RNA viruses and expressed viral genes, but the presence of viral RNA alone does not necessarily indicate an active exogenous infection. Viral sequences integrated into animal genomes during previous infections can remain transcriptionally active and therefore appear in RNA-seq datasets [18,19]. Comparing transcript-derived viral candidates against genomic EVE loci is therefore important for distinguishing these sources of viral signal [39].

I identified 24,796 candidate EVE loci across the host and symbiont genomes, of which 24,235 (97.74%) overlapped with repetitive genomic regions. Among the 2,714 Tier 3 viral-like transcripts, 2,468 mapped to the host or symbiont genomes. Of these mapped transcripts, 1,518 (61.5%) overlapped with candidate EVE loci and were therefore classified as transcribed EVEs, while the remaining 950 did not overlap with identified EVE loci and were retained as stronger candidates for exogenous viral sequences. Thus, without incorporating genomic information, a substantial portion of the Tier 3 transcriptional signal could have been interpreted as evidence of exogenous viruses rather than transcriptionally active genomic elements.

The *L. phyllopus*-associated virome could also have implications for host-microbial interactions. Insect-specific viruses can alter host physiology and immunity, and in mosquitoes, insect-associated viruses have been shown to influence infection with other viruses and alter vector competency [40,41]. Although our data does not provide evidence that viruses associated with *L. phyllopus* influence pathogen transmission, they provide a foundation for asking whether resident viruses could alter the ability of heteropteran hosts to acquire or transmit phytopathogens. Heteropterans are generally less prominent vectors of plant viruses than other suborders of Hemiptera, however, coreids like *Leptoglossus* spp. are agriculturally important pests and can vector other phytopathogenic organisms such as bacteria, fungi, and protozoans [42,43]. Future experiments will be needed to

determine whether persistent viral infections influence host susceptibility to microbial pathogens, gut colonization, or pathogen transmission.

Symbiont- and environment-associated Viral Elements. The taxonomic diversity recovered among the highest-confidence contigs emphasizes that the gut-associated virome represents multiple biological sources. Tier 1 sequences included putative RNA viruses, multiple sequences assigned to Caudoviricetes, and sequences with similarity to viral groups commonly associated with plants. Similar mixtures have been reported from other insect metatranscriptomes, where recovered viral sequences may originate from the insect host, its microbial community, or organisms encountered through feeding [39,44]. In our survey, viral sequences assigned to the plant-associated taxa Solemoviridae and Tombusviridae may represent viral RNA acquired through feeding rather than viruses actively replicating within *L. phyllopus*. Likewise, bacteriophage sequences recovered from the gut cannot currently be assigned to *Caballeronia* or another bacterial host. Determining the biological host of these viruses will require additional experimental work.

The bacteriophage-like sequences recovered here may nevertheless be particularly relevant to the symbiosis between *L. phyllopus* and *Caballeronia*. Sequences assigned to Caudoviricetes were the most frequently recovered group in the high-confidence dataset, indicating that transcriptionally active phage or phage-like sequences are present within the gut-associated microbial community. *Caballeronia* is acquired from the environment by heteropteran hosts each generation and subsequently colonizes specialized crypts within the M4 region of the host's midgut [5,6]. Following colonization, the symbiont-containing region becomes physically isolated from much of the remainder of the gut during host development [45]. In this context, even phages initially associated with the broader gut microbiome could be biologically important if they are capable of infecting *Caballeronia*. Disruption of the established symbiont population could potentially affect symbiont abundance or function and, consequently, host performance.

Environmental stress may provide one context in which these phage-symbiont interactions become particularly important. The transcriptomic libraries analyzed here were originally generated to investigate host and symbiont responses across different thermal conditions, making temperature a biologically relevant stressor in this system. Prophage induction can occur in response to environmental stress, including elevated temperature. For example, heat treatment of *Candidatus* Liberibacter asiaticus triggered a shift from lysogenic to lytic phage activity and increased phage production [46]. Although our pooled virome analysis cannot determine whether phage activity differed among the original thermal treatments, and there is currently no evidence that the phage-like sequences identified here infect *Caballeronia*, these results raise an interesting possibility for future work. Experimental exposure of symbiotic insects or cultured *Caballeronia* isolates to thermal stress, combined with measurements of phage abundance, symbiont density, and host performance, could determine whether phage activation contributes to variation in symbiont stability. Such interactions would add an additional biological layer to insect-symbiont responses to environmental stress beyond the direct effects of temperature on the host and bacterium.

Concluding Remarks. This study establishes an initial reference for the viral diversity associated with *L. phyllopus* and its gut-associated microbial community. Similar baseline surveys have provided useful points of comparison for detecting temporal and geographic changes in insect-associated viral communities [44]. Future work can build on

this foundation to determine which viruses persist across *L. phyllopus* and related Coreidae populations, which are associated specifically with insects or their microbial symbionts, and how viral abundance or activity changes across biological and environmental conditions. More broadly, our results demonstrate that the viral landscape associated with an insect-bacterial symbiosis consists of multiple overlapping signals, including candidate exogenous viruses, bacteriophages, environmentally acquired viral sequences, and transcriptionally active endogenous viral elements. Distinguishing among these sources provides a foundation for moving beyond viral discovery toward understanding whether viruses contribute to variation in host physiology, symbiont stability, and insect-microbe interactions.

Acknowledgements

I would like to thank Nicole Gerardo for all her support and members from the Gerardo Lab for their feedback on earlier drafts. Support for this research was provided by a United States Department of Agriculture NIFA grant (2024-67012-43746) awarded to Patrick T. Stillson.

Data Accessibility

All data used in these analyses are provided in the methods sections above as accession numbers to the *L. phyllopus* and symbiont genomes, and transcript data.

Conflict of Interest

The author declares that there are no conflicts of interest.

Table Legends

Table 1. Summary of the 20 high-confidence viral contigs identified from the *Leptoglossus phyllopus* gut-associated metatranscriptome. Taxonomic assignments and protein annotations are based on the top DIAMOND BLASTx hits against the NCBI non-redundant protein database, and transcript abundance is reported as transcripts per million (TPM). Complete annotations are provided in Table S1.

Figure Legends

Figure 1. Workflow used to identify exogenous viruses, endogenous viral elements (EVEs), and transcribed EVEs from host and symbiont datasets. Viral discovery analyses spanned from high-confidence viral predictions and taxonomic assignments to exploratory hidden Markov model (HMM)-based detection of viral domains. Candidate EVEs identified in genomic assemblies were further evaluated for repeat association and transcriptional activity. Tiers reflect decreasing confidence and increasing exploratory scope for divergent viral genetic elements.

Figure 2. Characteristics of the 20 high-confidence viral contigs recovered from the *Leptoglossus phyllopus* gut-associated metatranscriptome. (A) Distribution of viral contigs among the highest assigned taxonomic groups. Bar length indicates the number of recovered viral contigs, and shading represents the summed transcripts per (TPM) for each taxonomic group. (B) Lengths of individual recovered viral contigs. (C) Transcript abundance of individual viral contigs, measured as transcripts per million.

Figure 3. Viral discovery across top two confidence tiers. (A) Comparison of contigs identified in Tiers 1 and 2 based on mean amino acid identity to the best database match, median alignment length, and the number of viral taxa represented. Distribution of the highest supported viral taxonomic assignments among (B) Tier 1 and (C) Tier 2. Taxonomic assignments represent the highest supported viral rank recovered from sequence similarity searches and therefore include multiple taxonomic levels depending on annotation confidence.

Figure 4. Identification and characterization of candidate EVEs in *L. phyllopus*. (A) The majority of candidate EVE loci identified in the genome overlapped with annotated repetitive elements. (B) Nearly all candidate EVEs were associated with the *L. phyllopus* genome, while relatively few EVEs were detected in the symbiont genomes of *Caballeronia* strains LZ003 and LZ019. (C) Classification of Tier 3 viral-like transcripts following genomic mapping. Mapped transcripts were classified as transcribed EVEs and likely exogenous viral candidates.

References

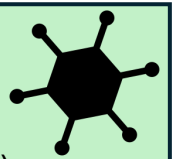
1. J. A. Noriega, J. Hortal, F. M. Azcárate, M. P. Berg, N. Bonada, M. J. Briones, I. Del Toro, D. Goulson, S. Ibanez, and D. A. Landis, Basic and Applied Ecology **26**, 8 (2018).
2. C. W. Schaefer and A. R. Panizzi, editors , *Heteroptera of Economic Importance* (CRC Press, Boca Raton, FL, 2000).
3. P. L. Mitchell, Denisia 1117 (2006).
4. P. L. Mitchell, Heteroptera of Economic Importance 337 (2000).
5. M. S. Hunter, E. F. Umanzor, S. E. Kelly, S. M. Whitaker, and A. Ravenscraft, Appl. Environ. Microbiol. AEM.01778 (2022).
6. Y. Kikuchi, T. Hosokawa, and T. Fukatsu, ISME J **5**, 446 (2011).
7. T. S. Acevedo, G. P. Fricker, J. R. Garcia, T. Alcaide, A. Berasategui, K. S. Stoy, and N. M. Gerardo, Front. Microbiol. **12**, 719112 (2021).
8. Y. Kikuchi, T. Hosokawa, and T. Fukatsu, Appl. Environ. Microbiol. **77**, 4075 (2011).
9. T. Ohbayashi, R. Cossard, G. Lextrait, T. Hosokawa, V. Lesieur, K. Takeshita, K. Tago, P. Mergaert, and Y. Kikuchi, Microbes Environ. **37**, n/a (2022).
10. B. C. Bonning, in *Insect Molecular Virology: Advances and Emerging Trends* (Caister Academic Press, 2019).
11. K. A. Coffman, A. N. Kauwe, N. E. Gillette, G. R. Burke, and S. M. Geib, Molecular Ecology **33**, e17485 (2024).
12. J. M. Kirsch, R. S. Brzozowski, D. Faith, J. L. Round, P. R. Secor, and B. A. Duerkop, Annu. Rev. Virol. **8**, 95 (2021).
13. M. J. Roossinck and E. R. Bazán, Annu. Rev. Virol. **4**, 123 (2017).
14. L. Giovannini, G. Mazza, W. Chitarra, G. Sabbatini-Peverieri, C. Sonnati, P. F. Roversi, and L. Nerva, Virus Research **316**, 198802 (2022).
15. C. Guo, Z. Ye, B. Hu, S. Shan, J. Chen, Z. Sun, J. Li, and Z. Wei, Viruses **14**, 2500 (2022).
16. S. Liu, Y. Chen, and B. C. Bonning, Current Opinion in Insect Science **8**, 54 (2015).
17. Y.-H. Qi, Z.-X. Ye, C.-X. Zhang, J.-P. Chen, and J.-M. Li, Computational and Structural Biotechnology Journal **21**, 4312 (2023).
18. A. Katzourakis and R. J. Gifford, PLoS Genet **6**, e1001191 (2010).
19. C. Feschotte and C. Gilbert, Nat Rev Genet **13**, 283 (2012).
20. D. J. Obbard, Current Opinion in Virology **31**, 17 (2018).
21. M. Martin, EMBnet. Journal 17.1 10 (2011).
22. D. Kim, B. Langmead, and S. L. Salzberg, Nat Methods **12**, 357 (2015).
23. D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, Bioinformatics **31**, 1674 (2015).
24. J. Guo, B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, T. O. Delmont, A. A. Pratama, M. C. Gazitúa, D. Vik, M. B. Sullivan, and S. Roux, Microbiome **9**, 37 (2021).
25. K. Kieft, Z. Zhou, and K. Anantharaman, Microbiome **8**, 90 (2020).
26. S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides, Nat Biotechnol **39**, 578 (2021).
27. B. Buchfink, C. Xie, and D. H. Huson, Nat Methods **12**, 59 (2015).
28. B. Bushnell, J. Rood, and E. Singer, PLoS ONE **12**, 1 (2017).
29. A. R. Quinlan and I. M. Hall, Bioinformatics **26**, 841 (2010).
30. P. Rice, I. Longden, and A. Bleasby, Trends in Genetics **16**, 276 (2000).
31. R. D. Finn, J. Clements, and S. R. Eddy, Nucleic Acids Research **39**, W29 (2011).
32. T. Bigot, S. Temmam, P. Pérot, and M. Eloit, F1000Research **8**, 530 (2020).
33. U. Singh and E. S. Wurtele, Bioinformatics **37**, 3019 (2021).
34. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, BMC Bioinformatics **10**, 421 (2009).
35. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit, Proc. Natl. Acad. Sci. U.S.A. **117**, 9451 (2020).
36. A. Smit, R. Hubley, and P. Green, (2013).
37. M. Shi, X.-D. Lin, J.-H. Tian, L.-J. Chen, X. Chen, C.-X. Li, X.-C. Qin, J. Li, J.-P. Cao, J.-S. Eden, J. Buchmann, W. Wang, J. Xu, E. C. Holmes, and Y.-Z. Zhang, Nature **540**, 539 (2016).
38. A. M. Ter Horst, J. C. Nigg, F. M. Dekker, and B. W. Falk, J Virol **93**, e02124 (2019).
39. P. Rozo-Lopez, W. Brewer, S. Käfer, M. M. Martin, and B. J. Parker, BMC Genomics **24**, 636 (2023).
40. B. Bolling, S. Weaver, R. Tesh, and N. Vasilakis, Viruses **7**, 4911 (2015).

41. R. P. Olmo, Y. M. H. Todjro, E. R. G. R. Aguiar, J. P. P. De Almeida, F. V. Ferreira, J. N. Armache, I. J. S. De Faria, A. G. A. Ferreira, S. C. G. Amadou, A. T. S. Silva, K. P. R. De Souza, A. P. P. Vilela, A. Babarit, C. H. Tan, M. Diallo, A. Gaye, C. Paupy, J. Obame-Nkoghe, T. M. Visser, C. J. M. Koenraadt, M. A. Wongsokarijo, A. L. C. Cruz, M. T. Prieto, M. C. P. Parra, M. L. Nogueira, V. Avelino-Silva, R. N. Mota, M. A. Z. Borges, B. P. Drumond, E. G. Kroon, M. Recker, L. Sedda, E. Marois, J.-L. Imler, and J. T. Marques, Nat Microbiol **8**, 135 (2023).
42. G. T. Broadhead, T. E. Cottrell, and J. J. Beck, J. Agric. Food Chem. acs.jafc.3c06763 (2024).
43. P. L. Mitchell, Neotrop. Entomol. **33**, 519 (2004).
44. M. Sadeghi, E. Altan, X. Deng, C. M. Barker, Y. Fang, L. L. Coffey, and E. Delwart, Virology **523**, 74 (2018).
45. Y. Kikuchi, T. Ohbayashi, S. Jang, and P. Mergaert, ISME J **14**, 1627 (2020).
46. F. Ding, V. Allen, W. Luo, S. Zhang, and Y. Duan, Hortic Res **5**, 30 (2018).

INPUT
Viral Transcriptome Assembly

Tier 1: High-confidence viral genomes (high confidence)

Methods:
- VirSorter2
- VIBRANT
- CheckV (quality filtering)



Output: confirmed viral genomes

Tier 2: Homology-supported viral contigs (high confidence)

Methods:
- DIAMOND BLASTx
- Retain viral taxa
- Filtering
- Remove Tier 1 contigs



Output: known viral homologs

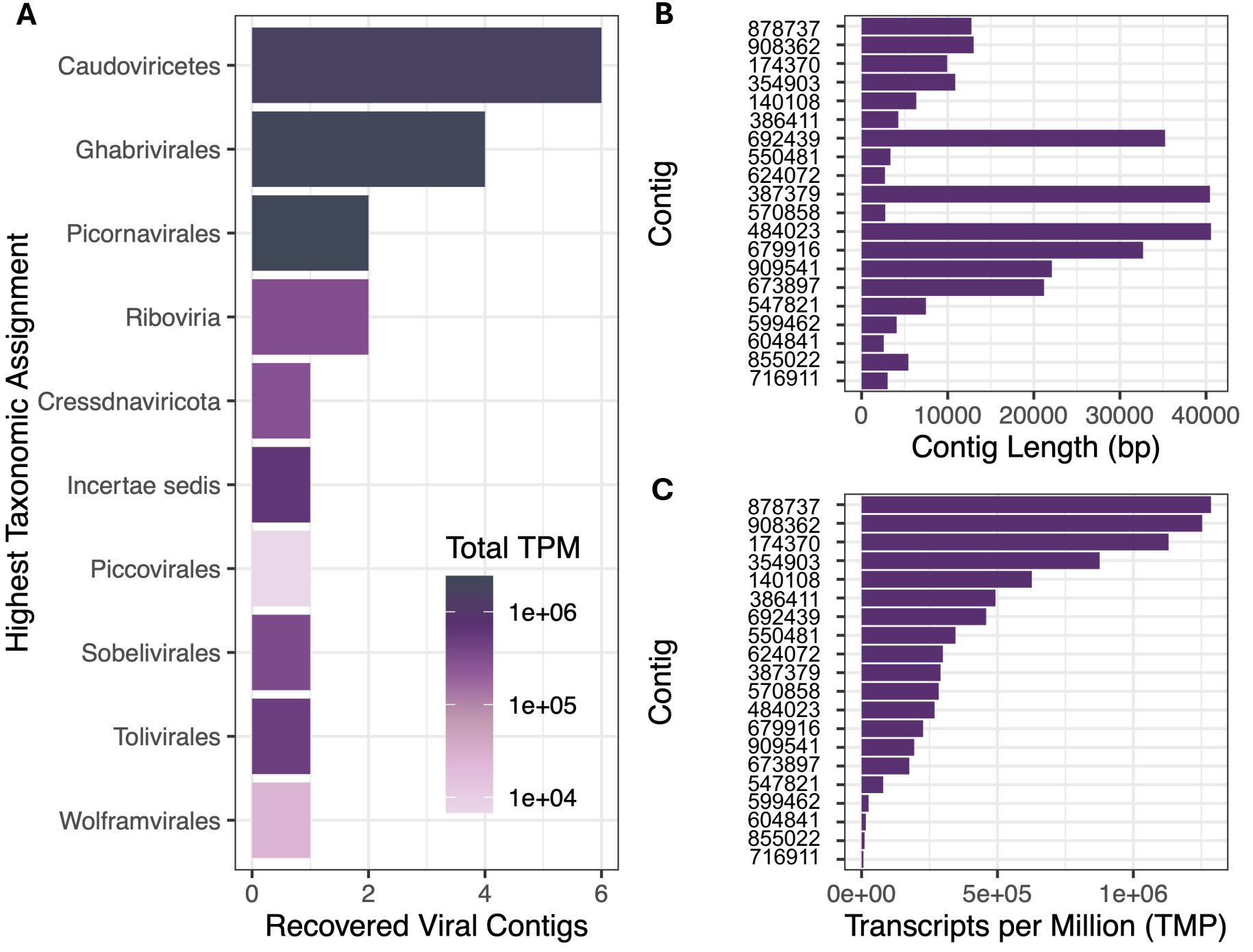
Tier 3: Viral-like domain-containing contigs (exploratory)

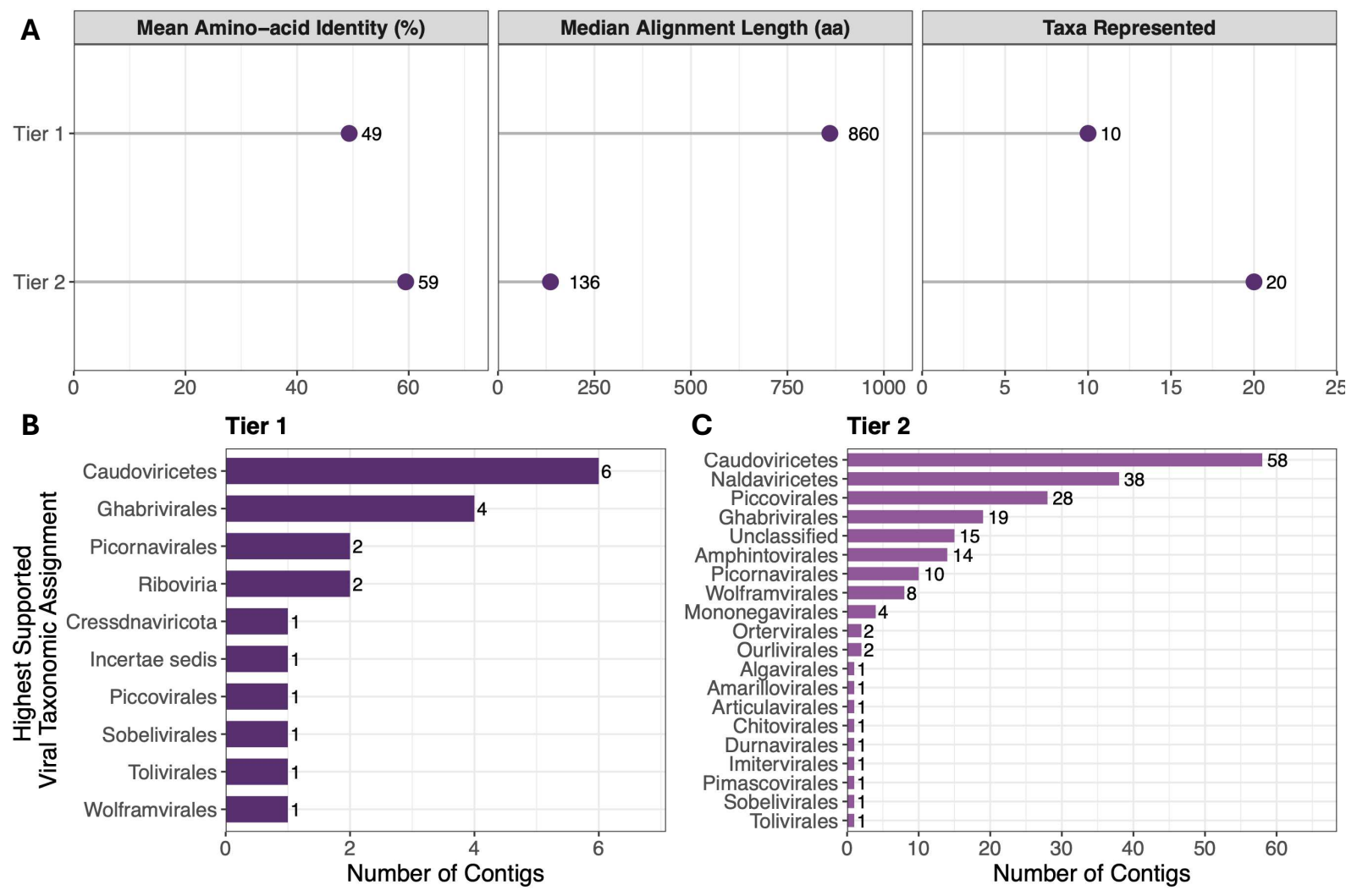
Methods:
- HMMER
- ORF translation
- Filtering (best domain per contig)
- Remove Tier 1 + Tier 2 contigs

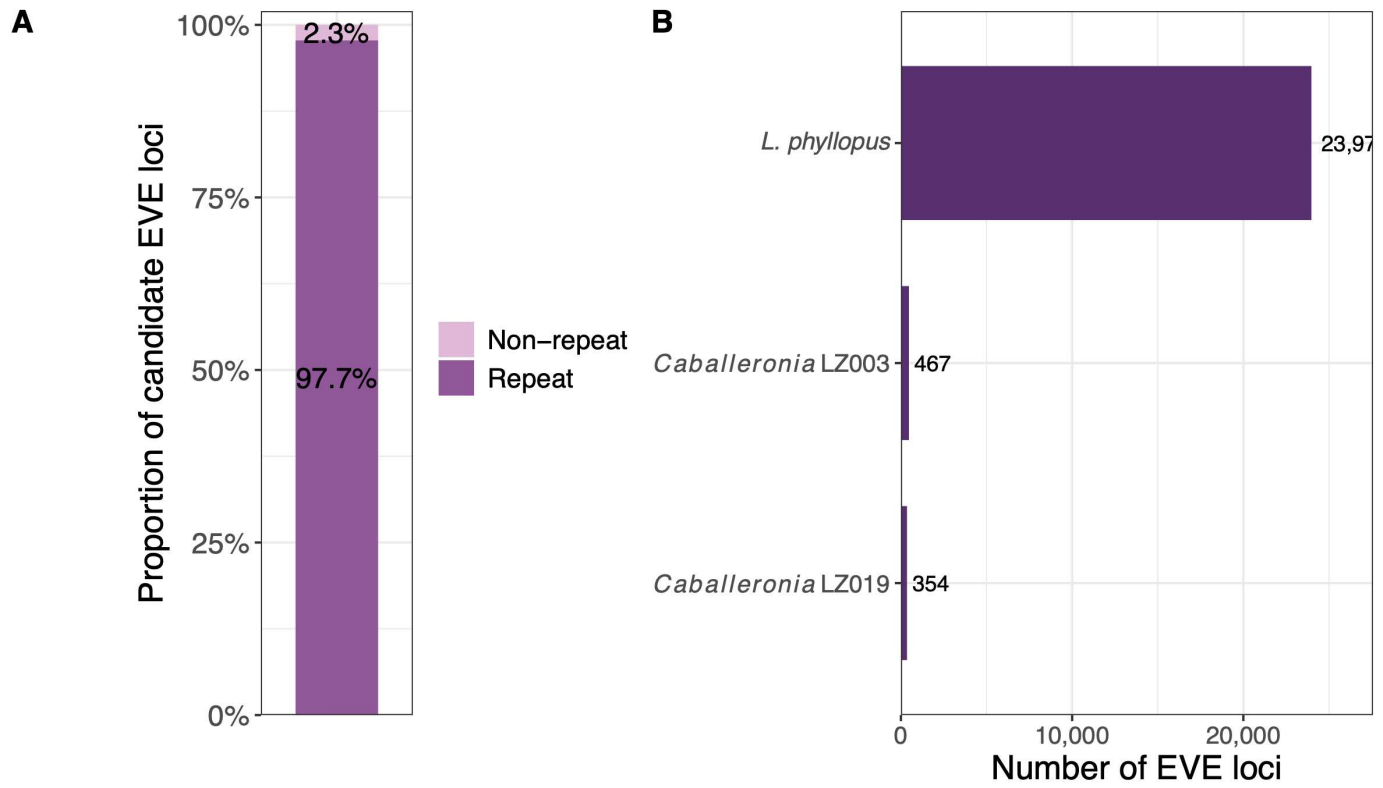
HMMER

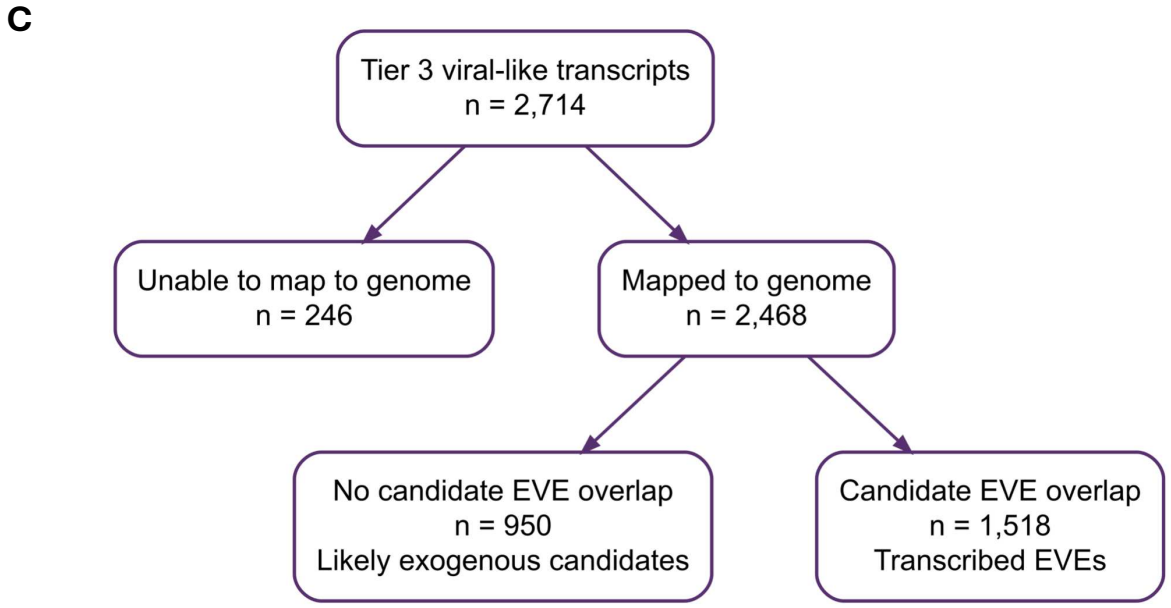
Output: viral-like protein domains

Final Output
Tier 1: 20 viral genomes
Tier 2: 207 contigs
Tier 3: 2,714 viral-like contigs


A
Highest Taxonomic Assignment
Caudoviricetes
Ghabrivirales
Picornavirales
Riboviria
Cressdnaviricota
Incertae sedis
Piccovirales
Sobelivirales
Tolivirales
Wolframvirales
Total TPM
1e+06
1e+05
1e+04
0
2
4
6
Recovered Viral Contigs
B
Contig
878737
908362
174370
354903
140108
386411
692439
550481
624072
387379
570858
484023
679916
909541
673897
547821
599462
604841
855022
716911
0
10000
20000
30000
40000
Contig Length (bp)
C
Contig
878737
908362
174370
354903
140108
386411
692439
550481
624072
387379
570858
484023
679916
909541
673897
547821
599462
604841
855022
716911
0e+00
5e+05
1e+06
Transcripts per Million (TMP)


A
Mean Amino−acid Identity (%)
Median Alignment Length (aa)
Taxa Represented
Tier 1
Tier 2
49
59
860
136
10
20
0
20
40
60
0
250
500
750
1000
0
5
10
15
20
25
B
Tier 1
Highest Supported Viral Taxonomic Assignment
Caudoviricetes
Ghabrivirales
Picornavirales
Riboviria
Cressdnaviricota
Incertae sedis
Piccovirales
Sobelivirales
Tolivirales
Wolframvirales
6
4
2
2
1
1
1
1
1
1
0
2
4
6
Number of Contigs
C
Tier 2
Caudoviricetes
Naldaviricetes
Piccovirales
Ghabrivirales
Unclassified
Amphintovirales
Picornavirales
Wolframvirales
Mononegavirales
Ortervirales
Ourlivirales
Algavirales
Amarillovirales
Articulavirales
Chitovirales
Durnavirales
Imitervirales
Pimascovirales
Sobelivirales
Tolivirales
58
38
28
19
15
14
10
8
4
2
2
1
1
1
1
1
1
1
1
1
0
10
20
30
40
50
60
Number of Contigs


A
100%
75%
50%
25%
0%
Proportion of candidate EVE loci
2.3%
97.7%
Non–repeat
Repeat
B
L. phyllopus
23,97
Caballeronia LZ003
467
Caballeronia LZ019
354
0
10,000
20,000
Number of EVE loci


C
Tier 3 viral-like transcripts
n = 2,714
Unable to map to genome
n = 246
Mapped to genome
n = 2,468
No candidate EVE overlap
n = 950
Likely exogenous candidates
Candidate EVE overlap
n = 1,518
Transcribed EVEs

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- SupplementalFigures.docx
- SupplementalTables.xlsx