Untangling an insect virome from its endogenous viral elements

Paula Rozo-Lopez,[1,*] William Brewer,[1] Simon Käfer,[2] McKayla M. Martin,[1] & Benjamin J. Parker [1,*]

[1] Department of Microbiology, University of Tennessee, Knoxville, TN 37916, USA

[2] Institut für Biologie und Umweltwissenschaften, Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany

* authors for correspondence: PRL: plopez2@utk.edu; BJP: bjp@utk.edu

ORCIDs: PRL - 0000-0001-9207-6579; SK - 0000-0003-3270-8348; MMM - 0000-0002-7471-6023; BJP - 0000-0002-0679-4732

ABSTRACT

Insects are an important reservoir of viral biodiversity, but the vast majority of viruses associated with insects have not been discovered. Recent studies have therefore employed high-throughput sequencing of RNA, which has led to rapid advances in our understanding of insect viral diversity. However, insect genomes frequently contain transcribed endogenous viral elements with significant homology to exogenous viruses, complicating the use of RNAseq for viral discovery. In this study, we use a multi-pronged sequencing approach to study the virome of an important agricultural pest and prolific vector of plant pathogens, the potato aphid *Macrosiphum euphorbiae*. We first used rRNA-reduced RNAseq to characterize the bacteria and viruses found in individual insects. We then characterized the frequency of a heritable Flavivirus and an Ambidensovirus in our population. We next generated a quality draft genome assembly for *M. euphorbiae* using Illumina-corrected Nanopore sequencing. This analysis showed that the Ambidensovirus, previously described from an RNAseq viral screen, is not a exogenous virus and instead is a transcribed endogenous viral element in the *M. euphorbiae* genome. Our study generates key insight into an important agricultural pest and highlights a widespread challenge for the study of viral diversity using RNAseq.

KEYWORDS

Viral discovery; RNAseq; insects; aphids; endogenous viral elements

INTRODUCTION

The last decade has transformed our understanding of the viral communities associated with insects, the most abundant and diversified animal group [1-4]. Insect viruses have been primarily studied in the context of vector-borne pathogens, which are transmitted horizontally between insect vectors and amplifying hosts and often have medical or agricultural relevance. Other viruses, however, only replicate within the insect and are maintained in natural populations through horizontal and/or vertical transmission. These insect-specific viruses have been shown to have important impacts on host biology [5-7], but much work remains to be done to describe insect-specific viral diversity and uncover the hidden role they play in insect phenotypes and evolution [8-10].

To address this gap, researchers have employed high-throughput approaches to viral discovery, including next-generation sequencing and analysis of RNA. Recent studies have used this approach to characterize and discover an enormous diversity of viruses [2, 11-15]. However, there are several serious limitations to this approach. For example, it is unclear from RNAseq data whether viral reads come from microbes infecting insect cells or if they are present from an organism ingested by the insect. Another potential challenge with using RNAseq for viral discovery is that insects often harbor fragments of viral sequences in their genomes. The endogenous viral elements (EVEs) described to date have homology with multiple clades of single- and double-stranded DNA and RNA viral families [16]. We have a limited understanding of the role EVEs are playing in insect biology, but transcriptionally active EVEs have been shown to play functional roles in regulating host genome stability and as an antiviral defense against exogenous viruses [17-19]. EVEs are remarkably common across insects [20], and thus EVEs could represent a widespread challenge facing the field. As such, studies are needed to uncover the contribution of EVEs to insect 'viromes'.

Aphids (Hemiptera: Aphidoidea) are hosts to diverse viruses, including plant pathogens with agricultural significance and insect-specific viruses [21, 22]. Recent studies have used metatranscriptome sequencing to describe viral diversity in aphids [23-27], and have described insect-specific DNA viruses in the family Parvoviridae (Ambidensovirus) and RNA viruses in the Bunyaviruses, Dicistroviruses, Flaviviruses, Iflaviruses, and Mesoviruses families [21]. The potato aphid *Macrosiphum euphorbiae* (Thomas, 1878) is an important cosmopolitan agricultural pest that infests tomatoes, potatoes, and other economically important crops [28]. *M. euphorbiae* is also an important vector of plant viruses (Families Bromoviridae, Closteroviridae, Geminiviridae, Potyviridae, and Solemoviridae) and was recently shown to host several insect-specific viruses belonging to the families Flaviviridae (Flavivirus) and Parvoviridae [24, 29, 30]. Despite *M. euphorbiae's* economic importance, no genomic resources are available outside the body and salivary gland transcriptomes [24, 31, 32].

The genomes of multiple aphid species have been shown to harbor EVEs that mediate growth, development, and wing plasticity [33-37]. In this study, we use next-generation sequencing and molecular techniques to show that aphid EVEs have led to the misidentification of aphid viruses from RNAseq data. First, we used RNAseq to characterize the microbial diversity of field-collected *M. euphorbiae* adults, and we found evidence of two insect-specific viruses infecting aphids collected from the field, including a Flavivirus and Ambidensovirus. Then, we generated a high-quality draft genome sequence of this species. Our genome showed that insect-specific Ambidensoviral hits correspond to transcriptionally active EVEs, indicating that a previously described virus is actually an endogenous viral element in the *M. euphorbiae* genome. Our results illustrate how careful analysis using multiple methods is needed to untangle insect viromes from EVEs, and this study furthers our understanding of the surprisingly widespread presence of densoviral EVEs in aphid genomes.

METHODS

Aphid collection: We collected asexual winged and wingless female *M. euphorbiae* adults from cultivated tomato plants (var Husky Cherry Red) in Knoxville, TN, USA, between April and June 2021 and 2022. We stored individual aphids in 1.5 mL Eppendorf tubes (Eppendorf, Hamburg, Germany) at -80°C until processing. To validate our ability to identify *M. euphorbiae* (NCBI TaxID: 13131), we used COI barcoding (LCO1490 5'-GGTCAACAAATCATAAAGATATTGG-3' and HCO2198 5'-TAAACTTCAGGGTGACCAAAAAATCA-3'), sanger sequencing, and comparisons of our COI sequences to the Barcode of Life Data System (https://www.boldsystems.org/) [38]. Our partial COI barcode sequence was uploaded to NCBI with accession number OQ588703.

Cultivation of *M. euphorbiae* strain Me57: To establish a colony of *M. euphorbiae* in the laboratory, we used a single asexual female collected in 2021. After colonization, we maintained this line on tomato plants (Husky Cherry Red) at 20°C 16L:8D. We screened the line for the seven species of facultative symbionts found in aphids using established PCR protocols [39, 40]. For this screen, we extracted DNA using 'Bender buffer' and ethanol precipitation as in previous studies [41, 42]. We then used PCR with species-specific primers [39, 43] to screen for *Hamiltonella defensa, Fukatsuia symbiotica* (X-type), *Regiella insecticola*, *Rickettsia* sp., *Ricketsiella* sp., *Serratia symbiotica*, and *Spiroplasma* sp. following the recommended thermal profiles (94°C for 2 min, 11 cycles of 94°C for 20 sec, 56°C (declining 1°C each cycle) for 50 sec, 72°C for 30 sec, 25 cycles of 94°C for 2 min, 45°C for 50 sec, 72°C for 2 min, and a final extension of 72°C for 5 min).

RNA extraction and sequencing: We homogenized individual aphids with a pestle in 500 µL of TRIzol (Invitrogen; Thermo Fisher Scientific, Inc., Waltham, MA, USA) and extracted total RNA using BCP (1-bromo-3-chloropropane; Life Technologies, Thermo Fisher Scientific, Inc., Waltham, MA, USA) with isopropanol precipitation. We used the Zymo RNA Clean & Concentrator kit (Zymo Genetics Inc., Seattle, WA, USA) to improve the purity and to remove gDNA using DNAse I. We then performed metatranscriptome Sequencing at Novogene (Novogene Corporation Inc., Sacramento, CA, USA). Library preparation was conducted using ribosomal RNA (rRNA) depletion by Illumina TruSeg Stranded Total RNA with Ribo-Zero Plus and NEBNext rRNA Depletion Kit (Zymo Genetics, Inc., Seattle, WA, USA). The libraries were sequenced to approximately 9 billion base pairs (bp) per sample with 150 bp paired-end reads on an Illumina NovaSeq platform. Raw reads were deposited into the NCBI Sequence Read Archive under BioProject ID PRJNA942253 with BioSample accessions SAMN33770905-SAMN33770908, and data accessions SRR23870213-SRR23870216.

Microbial analysis using CZID: We assessed the success of ribosomal reduction in the metatranscriptome libraries using riboPicker [44] and the reference database SILVA_138 [45] (supplementary file reads_report.csv). We then used the CZ ID platform pipeline V7.1 (https://czid.org) [46], a cloud-based, open-source bioinformatics platform designed to detect microbes from metagenomic data. We removed host-specific reads (STAR host subtraction) using the *Acyrthosiphon pisum* genome [47], trimmed adapters using Trimmomatic [48], removed low-quality reads with PriceSeqFilter [49], and aligned the remaining reads to the NCBI NT and NR databases using Minimap2 [50] and Diamond [51]. In parallel, short reads were *de* *novo* assembled using SPADES [52] and mapped back to the resulting contigs using bowtie2 [53] to identify the contig to which each raw read belongs. We used the CZ ID water background model, which evaluates the significance (z-scores) of relative abundance estimates for microbial taxa in each sample. Potential bacterial reads were distinguished from contaminating environmental sequences by establishing z-score metrics ≥10, alignment length over 50 matching nucleotides (NT L ≥50), and a minimum of five reads per million aligning to the reference protein database (NR rPM ≥ 5). Potential viruses were established by z-score metrics of ≥1, NT L ≥50, and NR rPM ≥ 5 [46, 54, 55]. Bacterial and viral hits were confirmed with BLASTX and BLASTN manual searches. Only annotated non-host hits with revised Taxonomy IDs and BLAST-based match refinement were used for further analysis. The “Macrosiphum euphorbiae” project is viewable and searchable to anyone in CZ ID.

Densoviral analysis using *de novo* assembly and Travis: We conducted an additional screening and viral genome assembly of potential Ambidensoviruses using *de novo* transcriptome assemblies obtained as follows. We used Trimmomatic v.0.39 [48] to trim the sequence adapters and filtered low-quality/complexity reads, and assessed for post-trimming quality using FastQC v.0.11.9 [56]. Then, we used Trinity v.2.14 [57] to *de novo* assemble the remaining reads. We used TRAVIS (v.20221029, https://github.com/kaefers/travis) to scan the assembled transcriptomes for Densovirus-like sequences. We built the reference database according to the currently accepted Densovirinae (ICTV, 29. Oct 2022, see supplementary file parvoviridae_reference_library.csv), extracted open reading frames between 100 and 2000 amino acids from the assembled transcriptomes, and screened using HMMER v3.3.1 [58], MMSeqs2 [59], BLASTP v2.12.0 [60], and Diamond v2.0.15 [51]. We set the e-value cutoff at 1 $\times 10^{-6}$, where applicable. All hits were again searched with Diamond against the non-redundant protein database (NCBI, downloaded on 29 Oct 2022).

MeV-1 genome analysis: We used the CZ ID viral consensus genomes pipeline to build a consensus genome from the sample with MeV-1 present at high levels. In short, contigs were aligned to the reference MeV-1 genome (NCBI Entry KT309079.1) using minimap2 [50] and then trimmed using TrimGalore (Phred score <20) [61]. The consensus genome was generated with iVar consensus using a depth of five or more reads [62].

MeV-1, MeV-2, and *Hamiltonella defensa* screening: Like all aphids, *M. euphorbiae* hosts an obligate heritable bacterial symbiont called *Buchnera aphidicola* that synthesizes amino acids missing from the aphid’s diet of plant phloem, and can also harbor several other facultative symbiotic bacteria (listed above) [43]. To screen for these microbes, we used 1 µg of total RNA extracted (as above) from each of the 23 adults collected during 2022 for cDNA synthesis with iScript cDNA synthesis kit (Bio-Rad Laboratories, Inc., Hercules, CA, USA). To screen for the Flavivirus Macrosiphum euphorbiae virus 1 (MeV-1), we used 100 ng of cDNA, the primers MevirF1 (5'-GTACACTTGCCTTACCTTACTGT-3') and MevirR1b (5'-AACACGGGTCACGACCTTAG-3'), and the PCR conditions previously described [30]. To screen for the Ambidensovirus Macrosiphum euphorbiae virus 2 (MeV-2), we used 100 ng of cDNA, the MeV2-F (5'-CCGGATGACAAATCCCACGA-3') and MeV2-R (5'-AATAGGCGCAGAGATGGACG-3') primers, and the recommended PCR conditions [24]. In addition, we extracted DNA from colonized Me57 aphids (as above) and used 40 ng of genomic DNA to screen for MeV-2. The aphid Glyceraldehyde 3-phosphate dehydrogenase (G3PDH) was used as internal control (primers G3PDH_F (5'-CGGGAATTTCATTGAACGAC-3') and G3PDH_R (5'- TCCACAACACGGTTGGAGTA-3') [35]).

We used 200 ng of the cDNA previously synthesized for MeV-1 and MeV-2 screening and the protocols for *Hamiltonella defensa* PCR screening (as described above) to evaluate the proportion of field-collected aphids harboring this bacterial symbiont (supplementary file samples_metadata.csv). Furthermore, we used a non-parametric (Spearman) correlation to investigate the potential interaction between *Hamiltonella* and MeV-1.

DNA extraction and sequencing: We pooled seven genetically identical adult unwinged aphids from cultivated lab line Me57 and isolated genomic DNA (gDNA) using a phenol/chloroform extraction. We then sheared the gDNA to approximately 20kb fragments using Covaris G-tubes (Covaris LLC., Woburn, MA, USA) at 4200 RMP for 1 minute, followed by tube inversion. For library preparation, we used the NEB Next PPFE repair kit with Ultra II end prep reaction (New England Biolabs, Ipswich, MA, USA) under recommended conditions and Nanopore ligation sequencing kit SQK-LSK110. For sequencing, we used a Nanopore R9.4.1 (FLO-MIN106D) flow cell and a MinION MIN-101B sequencing device (Oxford Nanopore Technologies, Oxford, UK). We ran the flow cell for 24 hours, followed by a wash with Flow Cell Wash Kit (EXP-WSH004); we then reloaded the flow cell with a second library prep and ran the sequencer for an additional 48 hours. We stopped the second sequencing run at 72 hours (~22 Gbps of sequencing). In addition, we performed an additional 5.3 Gb of 150 bp paired-end sequencing to polish the assembly on an Illumina NovaSeq platform. DNA was extracted as above, and library prep and sequencing were performed by Novogene Inc. Raw reads were filtered for low quality and adapter contamination by Novogene Inc.

***M. euphorbiae* whole genome assembly:** We used Guppy (Oxford Nanopore Technologies) for base-calling and quality trimming raw reads. For the removal of *Buchnera* reads, we used minimap2 v.2.24 [50] in conjunction with SAMtools v.1.15.1 [63] to map our reads against the *Buchnera aphidicola* (strain *Macrosiphum euphorbiae*) genome (NCBI accession NZ_CP029205) and the corresponding plasmids (NCBI accession number NZ_CP029203 and NZ_CP029204). We only kept unmapped reads for aphid genome assembly. We assembled Nanopore reads using CANU v.2.0 [64] with an estimated genome size of 541 Mbp. We removed allelic variants from the assembly using the purge_haplotigs v.1.1.2 [65], first by mapping reads to the assembly using minimap2 v2.24-r1122 with Samtools v.1.15.1 and manually choosing cutoffs for haploid vs. diploid coverage based on a histogram plot (v -l 5 -m 27 -h 60), and then by purging duplicated contigs based on coverage level (-j 80 -s 50). For assembly polishing, we used the Illumina reads after quality assessment using FastQC V0.11.9 [56]. Then we used these reads to polish the purged assembly using Pilon v.1.24 with default parameters [66]. We used BlobTools2 [67] to identify remaining contaminating contigs. For this, we used blast results obtained from the BLASTN function against the NR database using blast plus v.2.12.0 [68], read coverage obtained by mapping the Illumina reads to the assembly using minimap2 v.2.24 [50], and GC content in this analysis. Based on these results, we removed all the short contigs with strong homology to the plant genus *Solanum* (which includes the tomato host plant species of *M. euphorbiae*) as we suspect these contigs were assembled from host plant contamination in the guts of sequenced aphids. We also removed two short contigs with homology to other bacterial contaminants such as *Escherichia coli* and *Pseudomonas* sp. Lastly, we removed a contig nearly identical to the pLeu plasmid found in *Buchnera aphidicola* and a small portion of two large contigs matching the *Buchnera* genome. The final annotation was assessed using BUSCO v.5.3.2 [69] with the MetaEuk gene predictor [70] implemented in galaxy.org, using the hemiptera_odb10 (2020-08-05) lineage dataset. The *M. euphorbiae* genome is available in NCBI with BioProject ID PRJNA942253 and BioSample SAMN33681650. The raw Nanopore (SRR23851809) and Illumina reads (SRR23919025) associated with the genome are available through the Sequence Read Archive, and the finished assembly is available with accession number JARHUA000000000.

Characterizing endogenous viral elements in the *M. euphorbiae* genome: DNA Illumina raw reads were used as input to the CZ ID platform pipeline V7.1 (https://czid.org) and a z-score metrics of ≥1 and NT L ≥50 as described above [46, 54]. Additionally, to screen for actively transcribed Ambidensovirus-like EVES in the *M. euphorbiae* genome, we used BLASTN searches using the seven viral hits provided in individual Trinity contigs flagged by TRAVIS (supplementary file contigs_TRAVIS.fasta) against the genome scaffolds. All non-redundant hits from these searches with E-values < 1.10^{-3} were extracted and used in further analyses [33].

RESULTS

Analysis of non-host sequences detected in single aphid: We used the pea aphid (*A. pisum*) genome to subtract host reads from our transcriptome data set. On average, 81.8% of the reads mapped to this host and were subtracted from further analysis (see supplementary file reads_report.csv). We then analyzed the remaining distribution of non-pea aphid reads, within a single *M. euphorbiae* aphid, as the overall proportion of reads assembled into contigs that could be assigned to bacterial, eukaryotic, and viral taxa (public project Macrosiphum euphorbiae at https://czid.org). Bacterial taxa dominated the microbial signature (Figure 1A), and as expected, the highest number of reads assembled into contigs matched the aphid obligate symbiont *Buchnera aphidicola* with over 45,000 reads per million aligning to the nucleotide database (NT rPM>45,000). Reads from an aphid facultative symbiont *Hamiltonella defensa,* were found in two samples (NT rPM>8,700). One sample (Me152) showed a strong signature of bacterial contaminants (*E. coli, Pseudomonas, Halomonas,* and *Agrobacterium*) commonly present in soil and plant surfaces.

In terms of eukaryotes (Figure 1B), we found hits to Solanaceae, which includes the host plant species of *M. euphorbiae,* and Brachonidae parasitoid wasps (Insecta: Hymenoptera) in two samples (NT rPM>18,000). *M. euphorbiae* is known to be parasitized by hymenopterous wasps belonging to the superfamilies Ichneumonoidea (Braconidae) and Chalcidoidea [71]. In addition, there were some *M. euphorbiae* species-specific reads remaining, which did not map to the pea aphid reference genome but showed some homology to other aphid species (Insecta: Hemiptera).

We detected the presence of two insect-specific viruses in our metatranscriptome data (Figure 1C). The highest number of hits matched a previously described insect-specific Flavivirus, called Macrosiphum euphorbiae virus 1 (MeV-1), which we detected in two samples (NT rPM = 234 and 4055 for Me112 and Me202, respectively). We also detected viral hits to an insect-specific Ambidensovirus (Me202 and Me152; NT rPM>60). Other viral reads in our samples included a Bracovirus in one of the samples that was parasitized with the Brachonidae wasp (Me202; NT rPM=1) and a Tombusvirus (Me152; NT rPM=2.9), a family of plant pathogenic viruses with a single-stranded positive-sense RNA genome. Lastly, we detected two phage genera, the *Hamitonella*-specific phage APSE (NT rPM>310) in the same samples found positive for this symbiont (Me112 and Me202) and Acinetobacter phage (NT rPM 0.5-18), a bacteriophage largely prevalent in the environment [72].

Figure 1. Details of the per aphid breakdown of non-host reads aligning to specific bacteria (**Figure 1a**), eukaryotic (**Figure 1b**), and viral (**Figure 1c**) taxa. Reads per million aligning to the nucleotide database (NT rPM) used as the quantitative metric in the heatmaps (see supplementary files heatmap_metrics.csv for metric details).

Comprehensive and quantitative analysis of insect-specific viruses: Using the CZ ID platform, we aligned five assembled contigs to the MeV-1 reference genome (NCBI accession KT309079) and found that they ranged between 85.8-97.2% nucleotide identity to the reference genome (Figure 2A). Our transcriptome retrieved 17,397 informative nucleotides allowing the assembly of a nearly complete genome for MeV-1. Our MeV-1 consensus genome has a coverage breadth of 79% and a coverage depth of 673.2x (NCBI accession OQ504571) (supplementary figure MeV1_coverage.tif). This single-stranded positive-sense RNA genome contains a single large ORF encoding a polyprotein of 7,333 amino acids, which is subsequently processed to generate structural and non-structural proteins [73]. Previous analysis indicated that the polyprotein motifs of MeV-1 helicase, methyltransferase, and RdRp are similar to domains in other *Flavivirus* (family Flaviviridae) [21, 30]. The characteristic secondary structures (RNA stem-loop) in *Flavivirus* genomes most likely contributed to the 5,283 missing bases in our MeV-1 consensus genome assembly [74].

Figure 2. Assembled *M. euphorbiae* transcriptome contigs aligning to previously described insect Flavivirus (**Figure 2a**) and Ambidensovirus (**Figure 2b**) (see supplementary files contigs_CZID.fasta and contigs_TRAVIS.fasta for sequence details).

In addition, using the CZ ID platform, we detected two contigs with 80% similarity to the non-structural protein 1 (NS1) of Dysaphis plantaginea Densovirus (DplDNV), single-stranded DNA insect-specific *Ambidensovirus* (family Parvoviridae) (supplementary file contigs_CZID.fasta). Due to the lack of a publicly available genome or partial viral sequences of Macrosiphum euphorbiae virus 2 (MeV-2), an Ambidensovirus previously described in the same aphid species [24], we were not able to explore the homology between both viruses. Therefore, we conducted a more extensive analysis of our RNAseq data using TRAVIS, a consistency-based virus detection pipeline for sensitive mass screening of transcriptomic data directed toward Parvoviridae proteins. While degrees of sequence identity between Densovirinae (a subfamily of viral species exclusively infecting arthropods) is very low, viral species often express NS1 and VP proteins, which are useful for parvovirus phylogenetic inferences [75]. We used the seven viral hits provided in individual Trinity contigs flagged by TRAVIS (supplementary file contigs_TRAVIS.fasta) to identify the ORF orientation and similarity and to construct a hypothetical genome assembly using DplDNV as the closest reference available (Figure 2B). We found three contigs with 68.8% to 81.3% similarity to the non-structural ORF1 (encoding for the NS1 protein) and two contigs with 68.8% to 86.2% similarity to the structural ORF (encoding for the VP protein). None of the assembled contigs showed similarity to DplDNV ORF2 (encoding for the NS2 protein). We only detected 70% similarity with the ORF2 of a distantly related Ambidensovirus (NCBI accession AMG693112), which genomic organization differs from previously reported aphid densoviruses [21]. Importantly, all densoviral NS1-like sequences also showed a high nucleotide similarity (72-85%) to the pea aphid APNS-2 (NCBI accession NC_042493.1 and NC_042494.1), an EVE that contributes to wing phenotypic plasticity in this species [35].

Insect-specific virus frequency in natural populations: To further investigate the infection frequency of MeV-1 and MeV-2 infections in natural populations, we used a PCR approach to screen 23 individual adult aphids collected during 2022 as well as aphids from our colonized *Macrosiphum* line (Me57). We found only 13 field-collected aphids positive for MeV-1 (54.2%) and 21 aphids (87.5%) positive for MeV-2, including the colonized individuals (Figure 3). We also tested the cDNA of field-collected aphids (previously screened for MeV-1) for the presence of *Hamiltonella defensa* and found that 54.2% of the aphids (n=13) were harboring this bacterial symbiont. We found that 41.7% of individuals (n=10) shared a co-infection between this Flavivirus and *Hamiltonella* (Figure 3), but this association was not statistically significant (p-value= 0.078; r= 0.375).

Figure 3. Frequency of Macrosiphum euphorbiae virus 1 (MeV-1), Macrosiphum euphorbiae virus 2 (MeV-2), and *Hamiltonella denfesa* infections in wild-collected (n=23) and colonized (n=1) aphids. All samples tested using cDNA for PCR screenings.

Genome sequencing for analysis of endogenous viral elements (EVEs): Our laboratory line (Me57) was found to be PCR positive for MeV-2, and we then used DNA sequences from a pooled sample of Me57 aphids to look for viral reads. We used the CZ ID platform as above to identify viral taxa using the Illumina DNA reads from our colonized Me57 aphid line. Surprisingly, we detected only a single contig with a low number of Ambidensoviral hits (NT rPM>0.329), which also showed 79.0% similarity to the DplDNV NS1 viral protein and 84.34% similarity to an uncharacterized genomic transcript in pea aphids (NCBI accession XM_029492170.1). Since both of our transcriptome and genomic data were unable to recover a complete or near-to-complete Ambidensovirus genome, we then suspected that these viral reads could correspond instead to actively transcribed EVEs, as previously reported in other closely related aphid species [33, 35].

To determine with certainty whether the Ambidensovirus hits found in our transcriptome data corresponded to an actively transcribed EVE, we assembled the first *M. euphorbiae* genome publicly available. We obtained a total of 4,223,264 nanopore reads (at an average of 5.21kb) and 35,578,886 Illumina reads (PE 150bp) from sequencing. After assembly, haplotig purging, polishing, and manual removal of plant and bacterial contigs, our assembly contained 2,176 contigs with an N50 length of 665kb and a total length of 545.7 Mb (Figure 4A). *M. euphorbiae* has a similar GC content (29.96%; Figure 4B) to other sequenced aphids (e.g., *A. pisum* at 29.6%, *M. persicae* at 30.1%, and *A. glycines* at 27.8%) [76, 77]. The size of our assembly is close to a recent estimation of the *M. euphorbia* genome size based on flow cytometry which was estimated at 531.7 Mb [76]. Similarly, an analysis of single-copy orthologs showed our assembly contains 98.5% complete BUSCOs, with 94% present in single copies and 4.5% duplicated (Figure 4C). An additional 1.2% of BUSCOs are fragmented, and 0.3% are missing. Together these results suggest that this draft of the genome is highly complete.

Figure 4. M. *euphorbiae* genome assembly metrics (**Figure 4a**), GC content and coverage (**Figure 4b**), and BUSCO metrics (**Figure 4c**).

We used the genome as a reference to screen for the seven individual Trinity contigs flagged by TRAVIS as potential Ambidensovirus in our previous analysis. Initially, we selected hits with E-values < 1.10^{-3} [33]; however, most of the 3,044 hits represent shorter sequences rather than the actual transcript length (see supplementary file expressed_densoviral_EVEs.csv); therefore, we restricted the search to matches consistently to the entire length of each transcript and E-values=0 (Table 1). No full-length hits in the genome were found for the two largest viral contigs assembled from transcriptome data (contig3 and contig4); instead, the best hits for these two contigs corresponded to 16-17% of the total length. In insects, the EVE repertoire varies between distinct populations of a given species and, in some cases, even between individuals within the same population [78]. This phenomenon explains why all the field aphid samples (n=3) that tested negative for MeV-2 by PCR amplified a product of approximately 500 bp, which is about half of the expected size reported for the primers used. Given that the genome assemblies and RNAseq data sets were derived from different aphid strains, it is not surprising the wide range of partial-length Ambidensovirus hits obtained in our analysis. However, we are confident that five full-length viral transcripts are constitutively expressed from three regions of the *M. euphorbiae* genome (tig00030708_pilon, tig00029914_pilon, and tig00027226_pilon).

Table 1. List of Ambidensoviral transcripts and the corresponding integrations in *M. euphorbiae* genome.

Transcriptome contig	Trans-cript length	Percent-age of identical sites	Hit end	Hit start	Genome contig	Query end	Query start
Travis_contig1	783	96.70%	783	1	tig00030708_pilon	198038	197258
Travis_contig1	783	98.90%	1	783	tig00029914_pilon	60345	59559
Travis_contig2	466	96.20%	416	1	tig00030708_pilon	198433	198018
Travis_contig2	466	99.80%	1	466	tig00029914_pilon	59579	59114
Travis_contig3	2155	-	-	-	-	-	-
Travis_contig4	2878	-	-	-	-	-	-
Travis_contig5	1174	99.90%	1	1174	tig00030708_pilon	92758	91585
Travis_contig6	1040	99.80%	1	1040	tig00030708_pilon	85562	84525
Travis_contig7	635	100.00%	635	1	tig00030708_pilon	86191	85557
Travis_contig7	635	84.90%	635	1	tig00027226_pilon	138266	137632

Discussion

RNAseq is becoming an essential tool for virus discovery. Our study illustrates how endogenous viral elements in insect genomes can be an obstacle to using RNAseq for characterizing viral diversity in arthropods. We used rRNA-depleted RNAseq along with bioinformatic tools to characterize the virome of an important insect pest species, the potato aphid *Macrosiphum euphorbiae.* Our analysis found two insect-specific viruses from the families Flavivirus and Ambidensovirus described in previous RNAseq studies. However, by sequencing and assembling the genome of this insect using long-read sequencing, we found that the Ambidensovirus is a transcriptionally active EVE rather than an exogenous virus. Endogenous viral elements encoded in the host genome are abundant in arthropod genomes, and thus EVE sequences in RNAseq studies are an important consideration for future studies of viral diversity in arthropods.

Densoviral EVEs have been shown to be transcriptionally active in two other aphid species: *Myzus persicae* and *A. pisum*. In pea aphids, two copies of a transcribed densoviral non-structural protein (termed the “APNS” genes) were found to be upregulated in response to crowded conditions and to be functionally linked to the plastic production of wings [35]. These genes had close homology with the non-structural genes of Dysaphis plantaginea densovirus (DplDNV), which, when infecting rosy apple aphids, causes them to be winged [79], suggesting the function of these viral genes had been conserved after endogenization. The transcribed EVEs we found in *M. euphorbiae* have significant homology to the pea aphid APNS genes, and it seems likely that these genes may also be playing a role in wing plasticity in *M. euphorbiae* though additional data is needed.

Our study contributes to the growing list of sequenced aphid genomes, which together show that transcribed densoviral EVEs are common in this insect group [33, 35, 37, 80]. Most identified EVEs in insect genomes correspond to unclassified single-stranded RNA viruses and viruses belonging to the families Rhabdoviridae and Parvoviridae [78]. Unlike RNA viruses, which may produce abundant short mRNAs that favor virus endogenization [20], Parvoviruses undergo a double-stranded DNA intermediate during nuclear replication, which along with the endonuclease activity of NS1 and the eukaryote double-stranded break repair mechanism may largely favor endogenization of this virus family [81, 82]. Previous studies have estimated that around 10% of the parvoviruses described in animals are likely integrated into host genomes, but in most cases, the EVE status remains uncertain due to unavailable or incomplete genomes for those species in which transcriptome data is available [75]. Multiple recent studies have described the presence of “new” densoviruses in aphid’s transcriptome [23, 26, 83]; however, our combined transcriptomic and genomic analyses suggest that some of those viral transcripts may likely correspond to actively transcribed EVEs instead of heritable exogenous viruses infecting aphids at very high rates.

Last, our study shed light on the biology of MeV-1, an insect-specific Flavivirus (family Flaviviridae), previously characterized by RNAseq studies of *M. euphorbiae* populations collected in France [30]. We found that this virus, contrary to previous reports, is present in a North American population of *M. euphorbiae*, and we found that it is highly prevalent. By assembling the genome of MeV-1 from our RNAseq data, we found that our local population is infected with a potentially distinct viral strain from previous studies. No obvious infection symptoms or abnormal phenotypes were observed in MeV-1-infected *M. euphorbiae* adults, and future studies are needed to determine what phenotypic effects this virus has on its host. Other heritable viruses have been found to interact with the secondary symbiotic bacteria found in aphids [84-86] but we did not find significant patterns of co-infections with the bacterial symbiont *Hamiltonella defensa*.

EVEs are common in insect genomes, and our results highlight this widespread challenge in studying insect viromes. Our study further emphasizes how combining sequencing methodologies is necessary to overcome the potential pitfalls of only RNAseq-based viral discovery. Careful consideration of the biological characteristics and genome structure of viruses discovered through RNAseq is essential [87]. In aphids and other widely study systems, the development of cultured cell lines is also imperative to isolate viral species described by sequence-based methods, to characterize viral replication, and for use in large-scale virus production that will facilitate future investigation of the complex interaction of aphid viruses and their hosts [21].

The relatively high transcription level of some EVEs suggests that viral integration may have important biological implications for the fitness of aphids. Likewise, uncovering the phenotypic effects of accurately described insect-specific viruses may also show promising targets for alternative control strategies of agriculturally destructive organisms while providing important foundational resources in the study of host-virus dynamics. Research efforts need to be done on the evolutionary dynamics of heritable viruses to better understand how they are acting as hidden drivers of host phenotypes.

DATA ACCESSIBILITY: All data is available through CZ ID (Macrosiphum euphorbiae project), NCBI BioProject ID PRJNA942253, or included as a supplemental file.

AUTHORS’ CONTRIBUTIONS: PRL and BJP conceived of the project. PRL and BJP wrote the manuscript.
PRL, WB, SK, BJP carried out the data curation, bioinformatic analysis and software validation.
PRL, WB, MMM carried out the molecular work. BJP contributed to funding acquisition. All authors edited the manuscript and approved the final version for submission.

CONFLICT OF INTEREST: We declare we have no competing interests.

FUNDING: This work was funded by National Science Foundation (NSF) Grant IOS-2152954 to BJP. BJP is a Pew Scholar in the Biomedical Sciences, funded by the Pew Charitable Trusts.

ACKNOWLEDGMENTS: Thanks to Keertana Tallapragada, Georgina Aitolo, and Mariam Sirag for help with insect stock maintenance. The University of Tennessee Knoxville's Office of Innovative Technologies provided access to the ISAAC-NG computational cluster.

REFERENCES

[1] Koonin, E.V. & Dolja, V.V. 2018 Metaviromics: a tectonic shift in understanding virus evolution. *Virus Research* **246**, A1-A3. (doi:https://doi.org/10.1016/j.virusres.2018.02.001).

[2] Zhang, Y.-Z., Shi, M. & Holmes, E.C. 2018 Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**, 1168-1172. (doi:https://doi.org/10.1016/j.cell.2018.02.043).

[3] Greninger, A.L. 2018 A decade of RNA virus metagenomics is (not) enough. *Virus Research* **244**, 218-229. (doi:https://doi.org/10.1016/j.virusres.2017.10.014).

[4] Stork, N.E. 2018 How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology* **63**, 31-45. (doi:10.1146/annurev-ento-020117-043348).

[5] Coatsworth, H., Bozic, J., Carrillo, J., Buckner, E.A., Rivers, A.R., Dinglasan, R.R. & Mathias, D.K. 2022 Intrinsic variation in the vertically transmitted core virome of the mosquito Aedes aegypti. *Mol Ecol* **31**, 2545-2561. (doi:10.1111/mec.16412).

[6] Longdon, B. & Jiggins, F.M. 2012 Vertically transmitted viral endosymbionts of insects: do sigma viruses walk alone? *Proc Biol Sci* **279**, 3889-3898. (doi:10.1098/rspb.2012.1208).

[7] Longdon, B., Day, J.P., Schulz, N., Leftwich, P.T., de Jong, M.A., Breuker, C.J., Gibbs, M., Obbard, D.J., Wilfert, L., Smith, S.C., et al. 2017 Vertically transmitted rhabdoviruses are found across three insect families and have dynamic interactions with their hosts. *Proc Biol Sci* **284**. (doi:10.1098/rspb.2016.2381).

[8] González, R., Butković, A. & Elena, S.F. 2020 Chapter Three - From foes to friends: Viral infections expand the limits of host phenotypic plasticity. In *Advances in Virus Research* (eds. M. Kielian, T.C. Mettenleiter & M.J. Roossinck), pp. 85-121, Academic Press.

[9] Simmonds, P., Aiewsakun, P. & Katzourakis, A. 2019 Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology* **17**, 321-328. (doi:10.1038/s41579-018-0120-2).

[10] Mauck, K.E. 2016 Variation in virus effects on host plant phenotypes and insect vector behavior: what can it teach us about virus evolution? *Current Opinion in Virology* **21**, 114-123. (doi:https://doi.org/10.1016/j.coviro.2016.09.002).

[11] Bolling, B.G., Weaver, S.C., Tesh, R.B. & Vasilakis, N. 2015 Insect-Specific Virus Discovery: Significance for the Arbovirus Community. In *Viruses* (pp. 4911-4928.

[12] Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E.C. & Zhang, Y.-Z. 2015 Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4**, e05378. (doi:10.7554/eLife.05378).

[13] Käfer, S., Paraskevopoulou, S., Zirkel, F., Wieseke, N., Donath, A., Petersen, M., Jones, T.C., Liu, S., Zhou, X., Middendorf, M., et al. 2019 Re-assessing the diversity of negative strand RNA viruses in insects. *PLOS Pathogens* **15**, e1008224. (doi:10.1371/journal.ppat.1008224).

[14] Liu, S., Chen, Y. & Bonning, B.C. 2015 RNA virus discovery in insects. *Current Opinion in Insect Science* **8**, 54-61. (doi:https://doi.org/10.1016/j.cois.2014.12.005).

[15] Wu, H., Pang, R., Cheng, T., Xue, L., Zeng, H., Lei, T., Chen, M., Wu, S., Ding, Y., Zhang, J., et al. 2020 Abundant and Diverse RNA Viruses in Insects Revealed by RNA-Seq Analysis: Ecological and Evolutionary Implications. *mSystems* **5**, e00039-00020. (doi:10.1128/mSystems.00039-20).

[16] Gilbert, C. & Belliardo, C. 2022 The diversity of endogenous viral elements in insects. *Curr Opin Insect Sci* **49**, 48-55. (doi:10.1016/j.cois.2021.11.007).

[17] Veglia, A.J., Bistolas, K.S.I., Voolstra, C.R., Hume, B.C.C., Planes, S., Allemand, D., Boissin, E., Wincker, P., Poulain, J., Moulin, C., et al. 2022 Endogenous viral elements reveal associations between a non-retroviral RNA virus and symbiotic dinoflagellate genomes. *bioRxiv*, 2022.2004.2011.487905. (doi:10.1101/2022.04.11.487905).

[18] Blair, C.D., Olson, K.E. & Bonizzoni, M. 2020 The widespread occurrence and potential biological roles of endogenous viral elements in insect genomes. *Curr Issues Mol Biol* **34**, 13-30. (doi:10.21775/cimb.034.013).

[19] Frank, J.A. & Feschotte, C. 2017 Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol* **25**, 81-89. (doi:10.1016/j.coviro.2017.07.021).

[20] Holmes, E.C. 2011 The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368-377. (doi:10.1016/j.chom.2011.09.002).

[21] Guo, Y., Ji, N., Bai, L., Ma, J. & Li, Z. 2022 Aphid viruses: A brief view of a long history. *Frontiers in Insect Science* **2**. (doi:10.3389/finsc.2022.846716).

[22] Ray, S. & Casteel, C.L. 2022 Effector-mediated plant–virus–vector interactions. *The Plant Cell* **34**, 1514-1531. (doi:10.1093/plcell/koac058).

[23] Feng, Y., Krueger, E.N., Liu, S., Dorman, K., Bonning, B.C. & Miller, W.A. 2017 Discovery of Known and Novel Viral Genomes in Soybean Aphid by Deep Sequencing. *Phytobiomes Journal* **1**, 36-45. (doi:10.1094/PBIOMES-11-16-0013-R).

[24] Teixeira, M.A., Sela, N., Atamian, H.S., Bao, E., Chaudhary, R., MacWilliams, J., He, J., Mantelin, S., Girke, T. & Kaloshian, I. 2018 Sequence analysis of the potato aphid *Macrosiphum euphorbiae* transcriptome identified two new viruses. *PLOS ONE* **13**, e0193239. (doi:10.1371/journal.pone.0193239).

[25] Kondo, H., Fujita, M., Hisano, H., Hyodo, K., Andika, I.B. & Suzuki, N. 2020 Virome Analysis of Aphid Populations That Infest the Barley Field: The Discovery of Two Novel Groups of Nege/Kita-Like Viruses and Other Novel RNA Viruses. *Frontiers in Microbiology* **11**. (doi:10.3389/fmicb.2020.00509).

[26] Li, T., Li, H., Wu, Y., Li, S., Yuan, G. & Xu, P. 2022 Identification of a Novel Densovirus in Aphid, and Uncovering the Possible Antiviral Process During Its Infection. *Frontiers in Immunology* **13**. (doi:10.3389/fimmu.2022.905628).

[27] Wamonje, F.O., Michuki, G.N., Braidwood, L.A., Njuguna, J.N., Musembi Mutuku, J., Djikeng, A., Harvey, J.J.W. & Carr, J.P. 2017 Viral metagenomics of aphids present in bean and maize plots on mixed-use farms in Kenya reveals the presence of three dicistroviruses including a novel Big Sioux River virus-like dicistrovirus. *Virology Journal* **14**, 188. (doi:10.1186/s12985-017-0854-x).

[28] Blackman, R.L., Eastop, V.F. & Museum, N.H. 2000 *Aphids on the World's Crops: An Identification and Information Guide*, Wiley.

[29] Xu, Y. & Gray, S.M. 2020 Aphids and their transmitted potato viruses: A continuous challenges in potato crops. *Journal of Integrative Agriculture* **19**, 367-375. (doi:https://doi.org/10.1016/S2095-3119(19)62842-X).

[30] Teixeira, M., Sela, N., Ng, J., Casteel, C.L., Peng, H.-C., Bekal, S., Girke, T., Ghanim, M. & Kaloshian, I. 2016 A novel virus from Macrosiphum euphorbiae with similarities to members of the family Flaviviridae. *Journal of General Virology* **97**, 1261-1271. (doi:https://doi.org/10.1099/jgv.0.000414).

[31] Planelló, R., Llorente, L., Herrero, Ó., Novo, M., Blanco-Sánchez, L., Díaz-Pendón, J.A., Fernández-Muñoz, R., Ferrero, V. & de la Peña, E. 2022 Transcriptome analysis of aphids exposed to glandular trichomes in tomato reveals stress and starvation related responses. *Sci Rep* **12**, 20154. (doi:10.1038/s41598-022-24490-1).

[32] Atamian, H.S., Chaudhary, R., Cin, V.D., Bao, E., Girke, T. & Kaloshian, I. 2013 In planta expression or delivery of potato aphid Macrosiphum euphorbiae effectors Me10 and Me23 enhances aphid fecundity. *Mol Plant Microbe Interact* **26**, 67-74. (doi:10.1094/mpmi-06-12-0144-fi).

[33] Clavijo, G., van Munster, M., Monsion, B., Bochet, N. & Brault, V. 2016 Transcription of densovirus endogenous sequences in the Myzus persicae genome. *Journal of General Virology* **97**, 1000-1009. (doi:https://doi.org/10.1099/jgv.0.000396).

[34] Jayasinghe, W.H., Kim, H., Nakada, Y. & Masuta, C. 2021 A plant virus satellite RNA directly accelerates wing formation in its insect vector for spread. *Nature Communications* **12**, 7087. (doi:10.1038/s41467-021-27330-4).

[35] Parker, B.J. & Brisson, J.A. 2019 A laterally transferred viral gene modifies aphid wing plasticity. *Current Biology* **29**, 2098-2103.e2095. (doi:https://doi.org/10.1016/j.cub.2019.05.041).

[36] Shang, F., Niu, J., Ding, B.-Y., Zhang, W., Wei, D.-D., Wei, D., Jiang, H.-B. & Wang, J.-J. 2020 The miR-9b microRNA mediates dimorphism and development of wing in aphids. *Proceedings of the National Academy of Sciences* **117**, 8404-8409. (doi:10.1073/pnas.1919204117).

[37] Liu, S., Coates, B.S. & Bonning, B.C. 2020 Endogenous viral elements integrated into the genome of the soybean aphid, Aphis glycines. *Insect Biochem Mol Biol* **123**, 103405. (doi:10.1016/j.ibmb.2020.103405).

[38] Foottit, R.G., Maw, H.E., CD, V.O.N.D. & Hebert, P.D. 2008 Species identification of aphids (Insecta: Hemiptera: Aphididae) through DNA barcodes. *Mol Ecol Resour* **8**, 1189-1201. (doi:10.1111/j.1755-0998.2008.02297.x).

[39] Henry, L.M., Peccoud, J., Simon, J.C., Hadfield, J.D., Maiden, M.J., Ferrari, J. & Godfray, H.C. 2013 Horizontally transmitted symbionts and host colonization of ecological niches. *Curr Biol* **23**, 1713-1717. (doi:10.1016/j.cub.2013.07.029).

[40] McLean, A.H.C., Hrcek, J., Parker, B.J., Mathe-Hubert, H., Kaech, H., Paine, C. & Godfray, H.C.J. 2020 Multiple phenotypes conferred by a single insect symbiont are independent. *Proc Biol Sci* **287**, 20200562. (doi:10.1098/rspb.2020.0562).

[41] Goldstein, E.B., de Anda Acosta, Y., Henry, L.M. & Parker, B.J. 2022 Variation in density, immune gene suppression, and co-infection outcomes among strains of the aphid endosymbiont <em>Regiella insecticola</em>. *bioRxiv*, 2022.2008.2028.505589. (doi:10.1101/2022.08.28.505589).

[42] Bender, W., Spierer, P., Hogness, D.S. & Chambon, P. 1983 Chromosomal walking and jumping to isolate DNA from the Ace and rosy loci and the bithorax complex in Drosophila melanogaster. *Journal of Molecular Biology* **168**, 17-33. (doi:https://doi.org/10.1016/S0022-2836(83)80320-9).

[43] Henry, L.M., Maiden, M.C.J., Ferrari, J. & Godfray, H.C.J. 2015 Insect life history and the evolution of bacterial mutualism. *Ecology Letters* **18**, 516-525. (doi:https://doi.org/10.1111/ele.12425).

[44] Schmieder, R., Lim, Y.W. & Edwards, R. 2012 Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* **28**, 433-435. (doi:10.1093/bioinformatics/btr669).

[45] Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. & Glöckner, F.O. 2007 SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188-7196. (doi:10.1093/nar/gkm864).

[46] Kalantar, K.L., Carvalho, T., de Bourcy, C.F.A., Dimitrov, B., Dingle, G., Egger, R., Han, J., Holmes, O.B., Juan, Y.-F., King, R., et al. 2020 IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience* **9**. (doi:10.1093/gigascience/giaa111).

[47] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. 2012 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21. (doi:10.1093/bioinformatics/bts635).

[48] Bolger, A.M., Lohse, M. & Usadel, B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120. (doi:10.1093/bioinformatics/btu170).

[49] Ruby, J.G., Bellare, P. & Derisi, J.L. 2013 PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865-880. (doi:10.1534/g3.113.005967).

[50] Li, H. 2018 Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100. (doi:10.1093/bioinformatics/bty191).

[51] Buchfink, B., Reuter, K. & Drost, H.G. 2021 Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366-368. (doi:10.1038/s41592-021-01101-x).

[52] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477. (doi:10.1089/cmb.2012.0021).

[53] Langmead, B. & Salzberg, S.L. 2012 Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359. (doi:10.1038/nmeth.1923).

[54] Bohl, J.A., Lay, S., Chea, S., Ahyong, V., Parker, D.M., Gallagher, S., Fintzi, J., Man, S., Ponce, A., Sreng, S., et al. 2022 Discovering disease-causing pathogens in resource-scarce Southeast Asia using a global metagenomic pathogen monitoring system. *Proc Natl Acad Sci U S A* **119**, e2115285119. (doi:10.1073/pnas.2115285119).

[55] Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A.L., Li, L.M., Logan, P., Ratnasiri, K. & Retallack, H. 2021 Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *eLife* **10**, e68353. (doi:10.7554/eLife.68353).

[56] Andrews, S. 2010 FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[57] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652. (doi:10.1038/nbt.1883).

[58] Wheeler, T.J. & Eddy, S.R. 2013 nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487-2489. (doi:10.1093/bioinformatics/btt403).

[59] Steinegger, M. & Söding, J. 2017 MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028. (doi:10.1038/nbt.3988).

[60] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402. (doi:10.1093/nar/25.17.3389).

[61] Martin, M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3. (doi:10.14806/ej.17.1.200).

[62] Grubaugh, N.D., Gangavarapu, K., Quick, J., Matteson, N.L., De Jesus, J.G., Main, B.J., Tan, A.L., Paul, L.M., Brackney, D.E., Grewal, S., et al. 2019 An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology* **20**, 8. (doi:10.1186/s13059-018-1618-7).

[63] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079. (doi:10.1093/bioinformatics/btp352).

[64] Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736. (doi:10.1101/gr.215087.116).

[65] Roach, M.J., Schmidt, S.A. & Borneman, A.R. 2018 Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460. (doi:10.1186/s12859-018-2485-7).

[66] Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. 2014 Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**, e112963. (doi:10.1371/journal.pone.0112963).

[67] Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. 2020 BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics* **10**, 1361-1374. (doi:10.1534/g3.119.400908).

[68] Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. 2022 Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-d26. (doi:10.1093/nar/gkab1112).

[69] Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. & Zdobnov, E.M. 2021 BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647-4654. (doi:10.1093/molbev/msab199).

[70] Levy Karin, E., Mirdita, M. & Söding, J. 2020 MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48. (doi:10.1186/s40168-020-00808-x).

[71] Clarke, H.V. 2013 Genotypic and endosymbiont-mediated variation in parasitoid susceptibility and other fitness traits of the potato aphid, Macrosiphum euphorbiae. (

[72] Turner, D., Ackermann, H.W., Kropinski, A.M., Lavigne, R., Sutton, J.M. & Reynolds, D.M. 2017 Comparative Analysis of 37 Acinetobacter Bacteriophages. *Viruses* **10**. (doi:10.3390/v10010005).

[73] Blitvich, B.J. & Firth, A.E. 2015 Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. *Viruses* **7**, 1927-1959. (doi:10.3390/v7041927).

[74] Mazeaud, C., Freppel, W. & Chatel-Chaix, L. 2018 The Multiples Fates of the Flavivirus RNA Genome During Pathogenesis. *Frontiers in Genetics* **9**. (doi:10.3389/fgene.2018.00595).

[75] François, S., Filloux, D., Roumagnac, P., Bigot, D., Gayral, P., Martin, D.P., Froissart, R. & Ogliastro, M. 2016 Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Scientific Reports* **6**, 30880. (doi:10.1038/srep30880).

[76] Wenger, J.A., Cassone, B.J., Legeai, F., Johnston, J.S., Bansal, R., Yates, A.D., Coates, B.S., Pavinato, V.A. & Michel, A. 2017 Whole genome sequence of the soybean aphid, Aphis glycines. *Insect Biochem Mol Biol*. (doi:10.1016/j.ibmb.2017.01.005).

[77] International Aphid Genomics, C. 2010 Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol* **8**, e1000313. (doi:10.1371/journal.pbio.1000313).

[78] Horst, A.M.t., Nigg, J.C., Dekker, F.M. & Falk, B.W. 2019 Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. *Journal of Virology* **93**, e02124-02118. (doi:doi:10.1128/JVI.02124-18).

[79] Ryabov, E.V., Keane, G., Naish, N., Evered, C. & Winstanley, D. 2009 Densovirus induces winged morphs in asexual clones of the rosy apple aphid, *Dysaphis plantaginea*. *Proceedings of the National Academy of Sciences* **106**, 8465-8470. (doi:doi:10.1073/pnas.0901389106).

[80] Nigg, J.C., Kuo, Y.W. & Falk, B.W. 2020 Endogenous viral element-derived piwi-interacting RNAs (piRNAs) are not required for production of ping-pong-dependent piRNAs from Diaphorina citri Densovirus. *mBio* **11**. (doi:10.1128/mBio.02209-20).

[81] Kapoor, A., Simmonds, P. & Lipkin, W.I. 2010 Discovery and characterization of mammalian endogenous parvoviruses. *J Virol* **84**, 12628-12635. (doi:10.1128/jvi.01732-10).

[82] Liu, H., Fu, Y., Xie, J., Cheng, J., Ghabrial, S.A., Li, G., Peng, Y., Yi, X. & Jiang, D. 2011 Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *J Virol* **85**, 9863-9876. (doi:10.1128/jvi.00828-11).

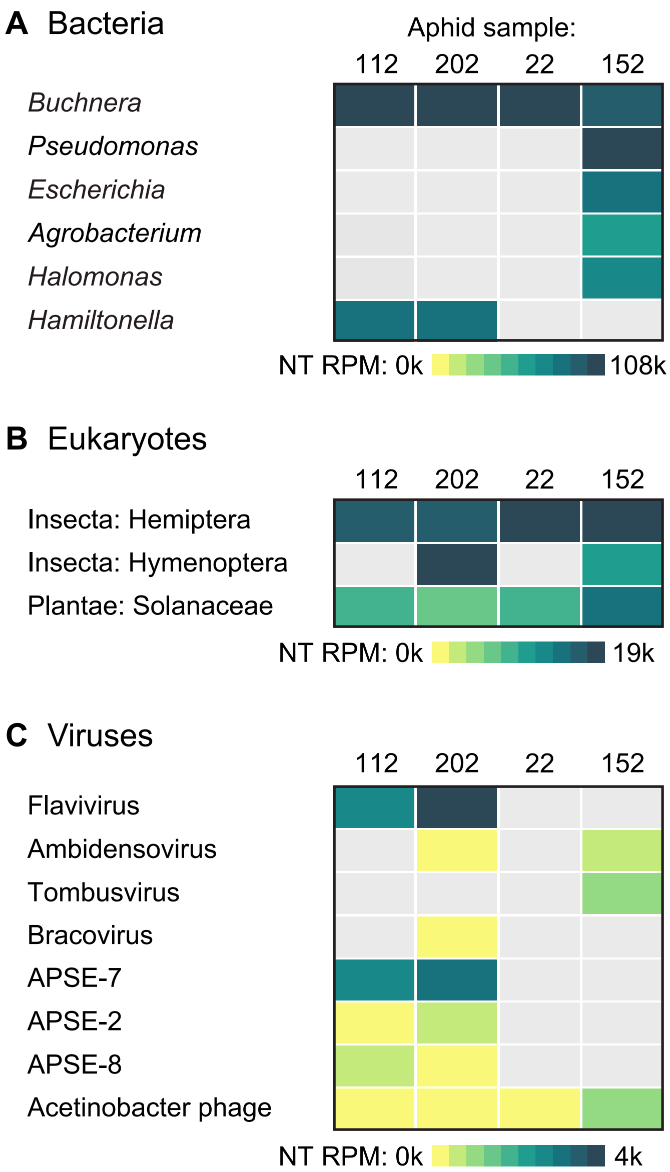
[83] Pinheiro, P.V., Wilson, J.R., Xu, Y., Zheng, Y., Rebelo, A.R., Fattah-Hosseini, S., Kruse, A., Dos Silva, R.S., Xu, Y., Kramer, M., et al. 2019 Plant viruses transmitted in two different modes produce differing effects on small RNA-mediated processes in their aphid vector. *Phytobiomes Journal* **3**, 71-81. (doi:10.1094/pbiomes-10-18-0045-r).

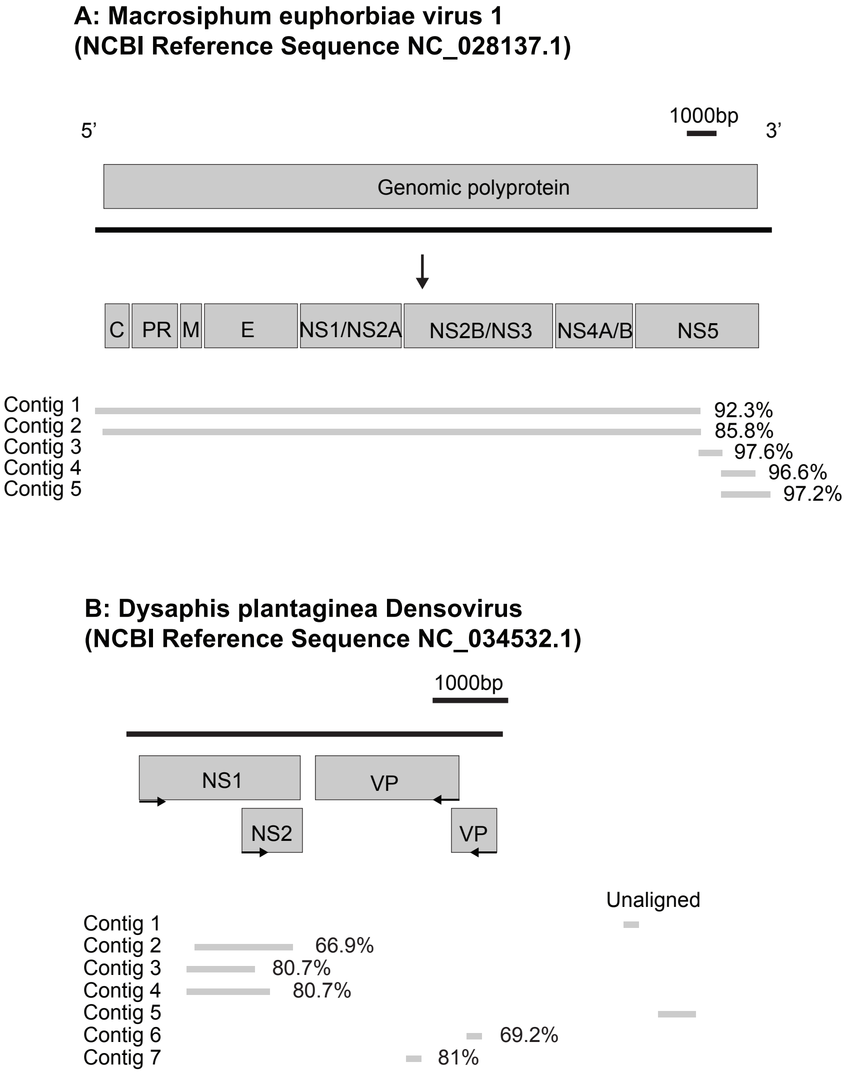
[84] Altinli, M., Schnettler, E. & Sicard, M. 2021 Symbiotic Interactions Between Mosquitoes and Mosquito Viruses. *Front Cell Infect Microbiol* **11**, 694020. (doi:10.3389/fcimb.2021.694020).

[85] Wu, W., Shan, H.W., Li, J.M., Zhang, C.X., Chen, J.P. & Mao, Q. 2022 Roles of Bacterial Symbionts in Transmission of Plant Virus by Hemipteran Vectors. *Front Microbiol* **13**, 805352. (doi:10.3389/fmicb.2022.805352).

[86] Jia, D., Mao, Q., Chen, Y., Liu, Y., Chen, Q., Wu, W., Zhang, X., Chen, H., Li, Y. & Wei, T. 2017 Insect symbiotic bacteria harbour viral pathogens for transovarial transmission. *Nature Microbiology* **2**, 17025. (doi:10.1038/nmicrobiol.2017.25).

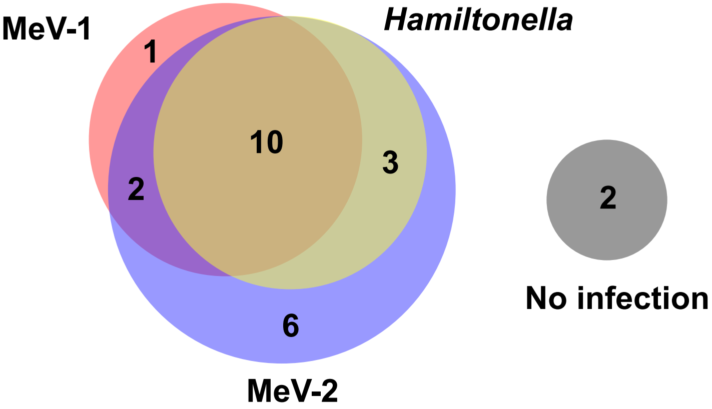
[87] Depledge, D.P., Mohr, I. & Wilson, A.C. 2019 Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *J Virol* **93**. (doi:10.1128/jvi.01342-18).

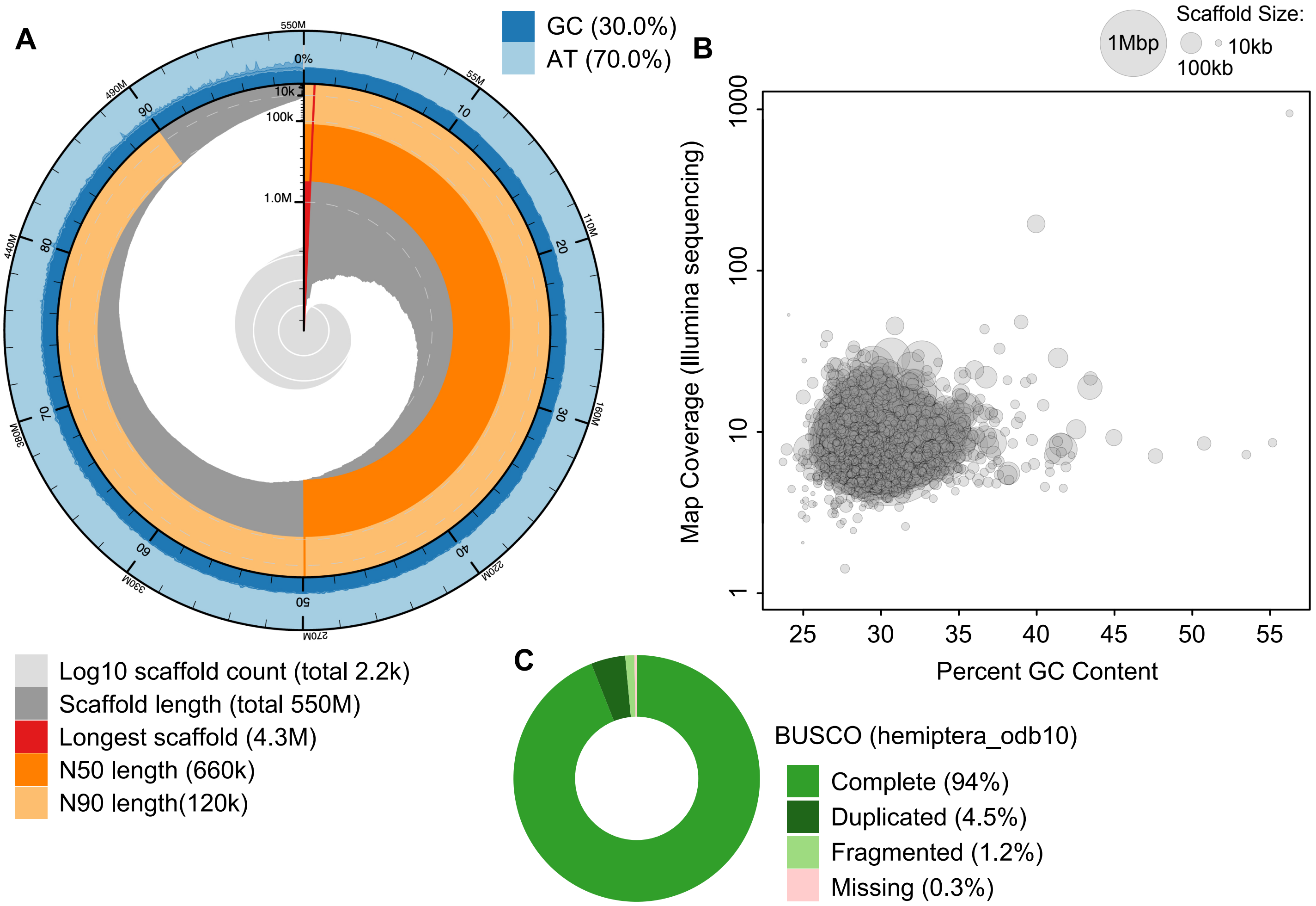

A Bacteria
Aphid sample:
112
202
22
152
Buchnera
Pseudomonas
Escherichia
Agrobacterium
Halomonas
Hamiltonella
NT RPM: 0k
108k
B Eukaryotes
112
202
22
152
Insecta: Hemiptera
Insecta: Hymenoptera
Plantae: Solanaceae
NT RPM: 0k
19k
C Viruses
112
202
22
152
Flavivirus
Ambidensovirus
Tombusvirus
Bracovirus
APSE-7
APSE-2
APSE-8
Acetinobacter phage
NT RPM: 0k
4k



A: Macrosiphum euphorbiae virus 1 (NCBI Reference Sequence NC_028137.1)

B: Dysaphis plantaginea Densovirus (NCBI Reference Sequence NC_034532.1)


MeV-1
Hamiltonella
1
10
3
2
2
6
No infection
MeV-2


A
550M
0%
55M
10
10k
100k
1.0M
110M
20
160M
30
220M
40
270M
50
330M
60
380M
70
440M
80
490M
90
GC (30.0%)
AT (70.0%)
Log10 scaffold count (total 2.2k)
Scaffold length (total 550M)
Longest scaffold (4.3M)
N50 length (660k)
N90 length(120k)
B
Scaffold Size:
1Mbp
100kb
10kb
Map Coverage (Illumina sequencing)
1000
100
10
1
25
30
35
40
45
50
55
Percent GC Content
C
BUSCO (hemiptera_odb10)
Complete (94%)
Duplicated (4.5%)
Fragmented (1.2%)
Missing (0.3%)