A tale of caution: How endogenous viral elements affect virus discovery in transcriptomic data

Nadja Brait, Thomas Hackl, Côme Morel, Antoni Exbrayat, Serafin Gutierrez, Sebastian Lequime

Virus Evolution, 2024, **10(1)**, vead088
DOI: https://doi.org/10.1093/ve/vead088
Advance Access Publication 28 December 2023
Research Article

A tale of caution: How endogenous viral elements affect virus discovery in transcriptomic data

Nadja Brait,[1] Thomas Hackl,[1,‡] Côme Morel,[2] Antoni Exbrayat,[2] Serafin Gutierrez,[2,§] and Sebastian Lequime[1,**,*]

[1]Cluster of Microbial Ecology, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen 9747 AG, The Netherlands and [2]ASTRE research unit, Cirad, INRAe, Université de Montpellier, Montpellier 34398, France
[‡]https://orcid.org/0000-0002-0022-320X
[§]https://orcid.org/0000-0002-5277-7239
[**]https://orcid.org/0000-0002-3140-0651
[*]Corresponding author: E-mail: s.j.j.lequime@rug.nl

Abstract

Large-scale metagenomic and -transcriptomic studies have revolutionized our understanding of viral diversity and abundance. In contrast, endogenous viral elements (EVEs), remnants of viral sequences integrated into host genomes, have received limited attention in the context of virus discovery, especially in RNA-Seq data. EVEs resemble their original viruses, a challenge that makes distinguishing between active infections and integrated remnants difficult, affecting virus classification and biases downstream analyses. Here, we systematically assess the effects of EVEs on a prototypical virus discovery pipeline, evaluate their impact on data integrity and classification accuracy, and provide some recommendations for better practices. We examined EVEs and exogenous viral sequences linked to *Orthomyxoviridae*, a diverse family of negative-sense segmented RNA viruses, in 13 genomic and 538 transcriptomic datasets of Culicinae mosquitoes. Our analysis revealed a substantial number of viral sequences in transcriptomic datasets. However, a significant portion appeared not to be exogenous viruses but transcripts derived from EVEs. Distinguishing between transcribed EVEs and exogenous virus sequences was especially difficult in samples with low viral abundance. For example, three transcribed EVEs showed full-length segments, devoid of frameshift and nonsense mutations, exhibiting sufficient mean read depths that qualify them as exogenous virus hits. Mapping reads on a host genome containing EVEs before assembly somewhat alleviated the EVE burden, but it led to a drastic reduction of viral hits and reduced quality of assemblies, especially in regions of the viral genome relatively similar to EVEs.
Our study highlights that our knowledge of the genetic diversity of viruses can be altered by the underestimated presence of EVEs in transcriptomic datasets, leading to false positives and altered or missing sequence information. Thus, recognizing and addressing the influence of EVEs in virus discovery pipelines will be key in enhancing our ability to capture the full spectrum of viral diversity.

Keywords: endogenous viral element; virus discovery; orthomyxovirus; mosquito.

Introduction

Decreasing costs of high throughput sequencing has allowed viral metagenomics to grow exponentially, with skyrocketing numbers of viral species deposited yearly (Wolf et al. 2020). Consequently, computationally generated genome assemblies have become sufficient evidence for their admission as *bona fide* viruses, no longer relying on cultivation and verification through phenotypic properties (Simmonds et al. 2017). A substantial subset of deposited genomes is derived from the re-analysis of genomic or transcriptomic datasets that were not always intended for virus discovery (e.g. Shi et al. 2016; Simmonds et al. 2017; Nayfach et al. 2021; Edgar et al. 2022; Johansen et al. 2022; Neri et al. 2022). While this opens up the spectrum of associated hosts and expands our knowledge of virus diversity, contaminating host nucleic acids are usually not removed before sequencing, taking away sequence capacity for viral reads from an already potentially low viral load in the samples (Prachayangprecha et al. 2014). The low abundance of viral reads combined with the presence of related species or populations in the same sample makes viral assembly challenging. Viruses with low and uneven read depth are known to produce fragmented assemblies or chimeric contigs if they are similar enough (García-López, Vázquez-Castellanos, and Moya 2015; Smits et al. 2015; Sutton et al. 2019). Viral reads with high sequence similarity may not only originate from similar viral populations in the host but could potentially come from the host genome itself in the form of prophages, provirus, or (non-) retroviral integration (Zhdanov 1975; Ackermann and DuBow 1987; Weiss 2006).

Endogenous viral elements (EVEs) are partial or complete insertions of viral genomes in the host's genome (Benveniste and Todaro 1974; Jaenisch 1976; Bejarano et al. 1996; Crochu et al. 2004). For such endogenization to occur, two essential steps need

to happen: (1) the production of dsDNA intermediates and (2) the integration into the host germline, probably through non-homologous recombinations or reverse transcription and integration driven by cellular retroelements, such as long interspersed nucleotide elements of the L1 family (Geuking et al. 2009; Belyi, Levine, and Skalka 2010; Holmes 2011; Tassetto et al. 2019; Wallau 2022). Most EVEs are of retroviral origin, as the integration of retroviruses is an obligatory part of their replication cycle (Herniou et al. 1998; Katzourakis, Rambaut, and Pybus 2005). Retroviral EVEs were first described in the 1970s (Benveniste and Todaro 1974), and despite non-retroviral EVEs being experimentally induced in the same decade (Zhdanov 1975), the first non-retroviral EVE in metazoans was only described 30 years later (Crochu et al. 2004). Since then, all Baltimore classes of viruses have been found to be integrated (Berns and Linden 1995; Geisler and Jarvis 2016; Horie et al. 2010; Katzourakis et al. 2007; H. Liu et al. 2011; W., 2012; Staginnus and Richertpoggeler 2006; Taylor, Leach, and Bruenn 2010).

Recent insertions and purifying selections within the host genome can lead to the emergence of EVEs similar to currently circulating viruses (Aiewsakun and Katzourakis 2015). The misclassification of EVE sequences as exogenous viruses is a particularly underestimated possibility in virus discovery studies based on (meta)-transcriptomics datasets. Indeed, depending on their integration sites, EVEs can exploit nearby regulatory elements and the host cell's transcriptional machinery, facilitating EVE transcription (Sofuku et al. 2018). While most described EVEs are highly mutated or fragmented, transcribing as non-coding RNAs, several have been discovered to encode intact open reading frames (ORFs) (Horie et al. 2010; Katzourakis, Gifford, and Malik 2010). In addition, during viral genome assembly, EVE-associated reads exhibiting high sequence similarity to exogenous viruses can inadvertently integrate into chimeric viral contigs or increase assembly fragmentation. Consequently, the accurate reconstruction and characterization of complete viral genomes becomes challenging. The common practice of host read removal in virus discovery pipelines, which could alleviate the burden of EVEs, can unintentionally exclude valuable viral-associated reads along with EVE sequences.

In this study, we aimed to examine the potential bias of transcribed EVEs in interpreting viral sequence assemblies in typical similarity-based virus discovery pipelines, such as NCBI BLAST searches for taxonomic classification. We focused our study on orthomyxoviruses in Culicinae mosquitoes. *Orthomyxoviridae* is a family of enveloped segmented negative-sense single-stranded RNA viruses that mainly infect vertebrates and arthropods. Since the first description of influenza A virus in 1933, orthomyxovirus discovery has been primarily driven by public health risk, comprising only a small group of mammal, bird, and tick-associated RNA viruses (Presti et al. 2009; Allison et al. 2015), but recent metatranscriptomics studies have detected a vast amount of novel orthomyxoviruses in invertebrates, greatly expanding the known host range and diversity of this family (Batson et al. 2021; Li et al. 2015; M. Shi et al. 2016). Complete orthomyxovirus genomes comprise between six and eight segments, with some segments only recently being identified, as can be seen for the genus *Quaranjavirus*, with two additional hypothetical proteins discovered (Batson et al. 2021). Despite the recent increase in orthomyxovirus diversity, most non-influenza genomes are not fully characterized, resulting in many species being taxonomically unclassified (Benson et al. 2013). Segmented viruses are often prone to loss of genomic information since most metagenomic studies predominantly screen only for the RNA-dependent RNA polymerase (RdRp). While non-segmented genomes can often be connected through linkage analyses, segmented viruses are more difficult to assemble, especially from diverse populations or low viral abundance (Krishnamurthy and Wang 2017).

Our analysis of publicly available genomic and transcriptomic datasets from Culicinae mosquitoes revealed the presence of orthomyxovirus-derived EVEs integrated into host genomes, particularly in *Aedes* species. We show that classifying contigs into EVE transcripts or exogenous-associated sequences for samples with low viral abundance is challenging and can potentially lead to misidentification. We also observed that EVEs in reference genomes for read removal can impact the downstream analysis of a typical virus discovery pipeline.

Material and methods

To assess the impact of EVEs on RNA virus detection, we first detected and characterized genomic EVEs, a prerequisite for identifying transcribed EVEs. We then compared a virus discovery pipeline's outcomes with and without EVEs to characterize their influence on data integrity and classification accuracy, particularly regarding read removal from host genomes and potential misclassification with similar exogenous viruses. The complete workflow of our pipeline is illustrated in Fig. 1.

Detection of EVEs in genomic datasets

Thirteen assemblies of whole genome sequence datasets of Culicinae mosquitoes were obtained from https://www.ncbi.nlm.nih.gov/Traces/wgs/ (Supplementary Table 1). The genomes used were assembled independently. However, for a minority of the studies, related genetic background strains were used (e.g. f.e Foshan strain, see Supplementary Table 1). Reference genomes, as well as transcript sequences from *Aedes albopictus* Foshan FPA, *Aedes aegypti* LVP_AGWG, and *Culex quinquefasciatus* Johannesburg, for host read removal were obtained from Vectorbase (Arensburger et al. 2010; Matthews et al. 2018; Palatini et al. 2020; Amos et al. 2022). Potential EVEs were screened and validated with a previously established in-house R script available on FigShare (https://figshare.com/projects/EVEs-bias-virus-discovery/185650) (Lequime and Lambrechts, 2017). To summarize, files were downloaded, merged, unzipped, and used as a database for a tblastn search (NCBI-blast-2.13.0) against curated orthomyxovirus reference sequences (accessions provided in Supplementary Table 2, sequences available on FigShare at https://figshare.com/projects/EVEs-bias-virus-discovery/185650) with an e-value threshold of 10^{-4}, as used previously (Katzourakis, Gifford, and Malik 2010; Lequime and Lambrechts 2017; Li et al. 2022). Putative EVE sequences were extracted, clustered, and merged as long as they overlapped or had a gap size of less than 100 nucleotides and were in the same orientation on the host contig. Sequence clusters ≥250 nt were used as the queries for a reciprocal blast search (NCBI-blast-2.13.0) against the NCBI nr and nt databases (both downloaded in April 2021) with an e-value cut-off of 10^{-4} to ensure viral origin. Contigs of interest were extracted, and genetic features of identified EVEs were checked with the NCBI Conserved Domain Database (Marchler-Bauer et al. 2015). Amino acid sequences were obtained by translating query sequences with diamond blastx (v2.0.13.151) and were manually checked afterward. To account for potential frameshift mutations, alignments produced by diamond BLAST were screened with *blast traceback operations (BTOP)*. EVEs within the reference sequences were cut out manually.

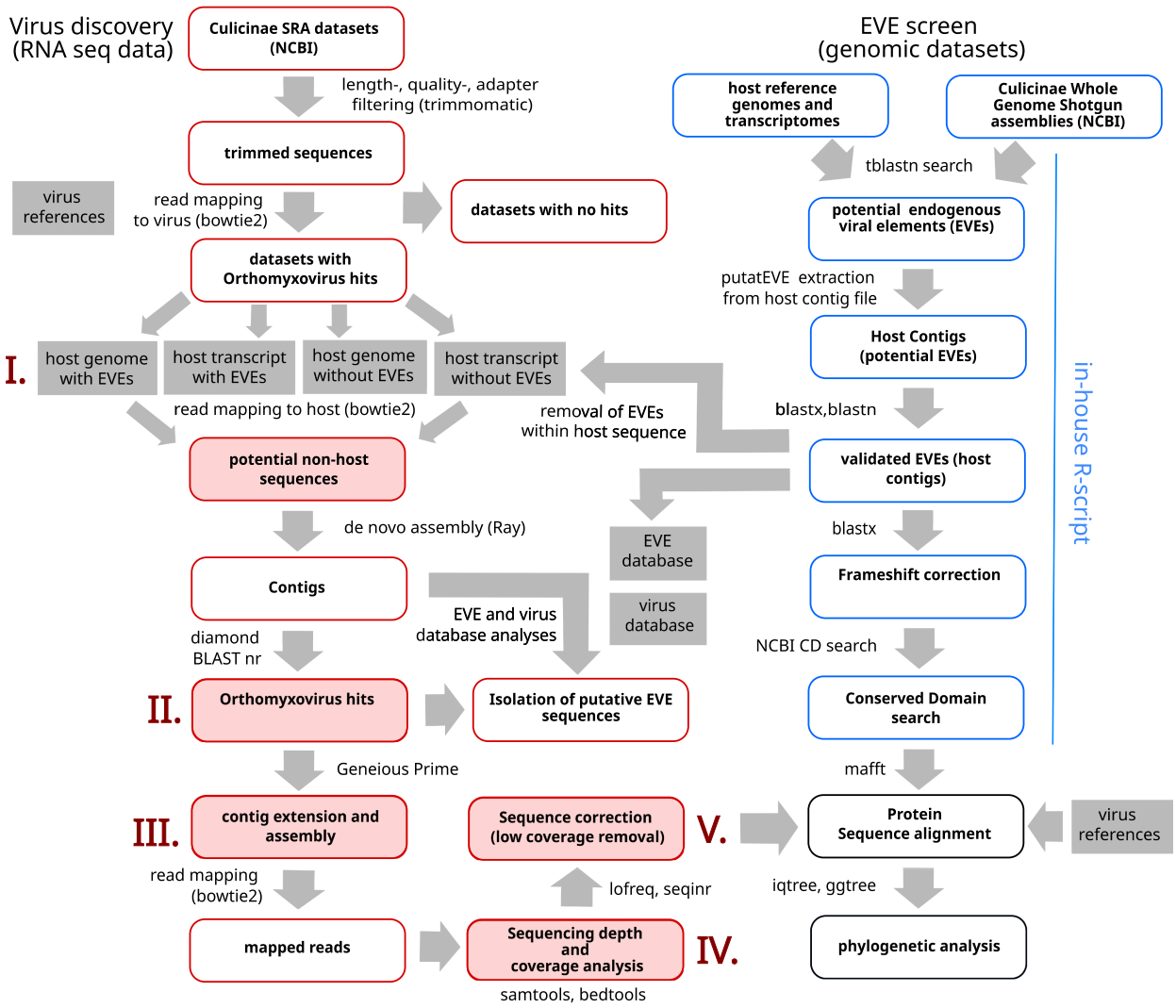

Figure 1. The influence of EVEs on the virus discovery process. Bioinformatics pipelines for transcriptomic virus discovery (tailored to exogenous orthomyxoviruses, framed in red) and genomic EVE detection (framed in blue). Pipeline steps potentially influenced by EVEs are shaded red: I. The presence of EVEs in host reference sequences may impact read filtering during host read removal. II. EVE sequences with high similarity to exogenous viruses could lead to false positive hits. III. EVE sequences with high similarity to exogenous viruses may erroneously assemble with exogenous viral sequences, resulting in chimeric assemblies. IV. Viral sequences accidentally filtered out during host read removal due to the presence of EVEs can lead to reduced sequencing depth and incomplete coverage. V. EVE sequences with high similarity to exogenous viruses might introduce SNPs in viral consensus sequences. The tools and packages utilized are indicated outside the respective boxes.

Collection of metadata in the Sequence Read Archive repository

To identify datasets of high-throughput RNA-Seq of Culicinae, we searched the Sequence Read Archive (SRA) at NCBI (Leinonen, Sugawara, and Shumway 2011). Datasets were sent to the SRA Run selector using the following search terms: 'Aedes' [All Fields] OR 'Culex' [All Fields] OR 'Culicinae' [All Fields]. We excluded genomic datasets, amplicons, target capture, heavily modified clones, small RNAs (e.g. microRNA), and datasets with unspecified experimental workflows (Supplementary Table 6). Origins of mosquito latitude and longitude positions were approximated from metadata and plotted using ggplot2 (v3.3.5 (Whickham 2016)) in the RStudio environment (v2021.09.1) (Fig. 3).

Sequence assembly using publicly available transcriptomic data

Fastq files of RNA-Seq datasets were obtained from the SRA database by efetch (v14.6) and pre-processed using fastq-dump (v2.10.9). Sequences were checked with fastqc (v0.11.9), and adapter, length, and quality trimmed with trimmomatic (v0.39). The trimmed reads were then passed to a relaxed Bowtie2 mapping with an orthomyxovirus-specific reference list (Supplementary Table 2). Host read removal was carried out by mapping the trimmed sequences of samples with putative positive hits relaxed against the above-mentioned host genomes and transcriptomes. This step was performed for the references with still integrated EVEs and those with initial EVE removal. Unmapped reads were *de novo* assembled with Ray (version 2.3.1). Assembled contigs were screened against the nr database with a full sensitivity for hits of >40 per cent identity from diamond-blastx (v2.0.13.151). Software, versions, and parameters can be found in the Supplementary Table under STAR methods.

Exploration of putative EVE-like sequences in transcriptomic data

To identify transcribed orthomyxovirus-derived EVEs in our RNA datasets, we used a similarity-based approach by comparing

contigs with positive orthomyxovirus hits to both genomic EVEs and publicly available exogenous sequences. We generated two custom databases of the existing orthomyxovirus reference list or the translated amino-acid protein sequences of previously detected genomic EVEs. The highest bit score hits of viral contigs against each database were merged into a table, and a Δbit-score value (bit score values of exogenous virus hits—bit score values of EVE hits) was determined. Bit scores to EVE hits were assigned a negative value, while hits corresponding to the non-EVE references remained positive. Pmax values (parallel maxima of two vectors, consisting of either the highest EVE or non-EVE bit score) and Δbit-scores were plotted with geom_jitter (ggplot2 v.3.3.5). Contigs with a Δbit-score lower or equal to −10 or with frameshift mutations were considered as putative EVE-like sequences and were further excluded for segment assemblies. This cut-off was set to allow the exclusion of contigs with a higher amino-acid similarity to EVE sequences, but would not exclude contigs with a similar low identity to both databases. This threshold was chosen empirically and should be taken cautiously.

Contig extensions and post-processing

To improve coverage and correct the newly obtained sequences, we extended the ends of each contig with partial segment lengths using corresponding complete genome segments and re-mapped all reads. Nucleotide reference sequences resulting in a positive hit for orthomyxoviruses were downloaded from NCBI GenBank with efetch (v14.6). Contigs were mapped to a non-redundant referencing sequence list corresponding to their best blastx hits in Geneious Prime (v.2022.1.1). Contigs spanning only parts of the orthomyxovirus segments were artificially extended to the full segment length by merging them with their reference sequence determined by their best blast hit. Non-host reads used for assembly were mapped back strictly ('unclipped' alignment) to the chimeric contigs using Bowtie2 (v2.3.5.1). The generated alignment file was converted, sorted, and indexed using Samtools (v1.10, htslib 1.10.2-3) with a mapping quality filter of MAPQ≥ 2. Coverage and sequencing depth were assessed using bedtools (v2.27.1). A sequence depth <3 was considered insufficient for subsequent analyses and nucleotides were replaced with ambiguous Ns. Single nucleotide variants were called using Lofreq* (version 2.1.2) and sequences were corrected when variant abundance reached >50 per cent. Coverage lengths and segment distribution per sample were illustrated with the ggplot2 function geom_tile (v3.3.5).

Phylogenetic analyses

Potential viral proteins, EVEs, and corresponding homologs of orthomyxovirus reference sequences were aligned with the Geneious implemented MAFFT aligner (v7.490) employing either the L-INS-i, FFT-NS-i or FFT-NS-2 algorithm, which was automatically adjusted to the input query (Katoh and Standley 2013). Because of the highly divergent nature of the sequence alignment, ambiguously aligned regions and gapped sites were pruned manually in Geneious. Sequences with more than 50 per cent unknown/missing residues (represented as X in amino-acid sequence) were excluded from the final alignment but can be found on FigShare (https://figshare.com/projects/EVEs-bias-virus-discovery/185650). Since most sequences were partial (either through sequence depth or as an EVE), no strict alignment lengths were applied. Phylogenetic trees were constructed using IQTREE (version 1.6.12) (Nguyen et al. 2015). The substitution models were selected based on the Bayesian information criterion provided by the IQTREE-implemented ModelFinder (Kalyaanamoorthy et al. 2017): Segment PB1: LG + F + I + G4, Segment PB2: LG + F + G4, Segment PA: LG + F + I + G4, Segment NP: VT + I + G4, Segment GP: VT + I + G4, Segment HP1: VT + G4, Segment HP2: FLU + F + G4, Segment HP3: VT + G4. Branch support values were measured as ultrafast bootstrap by UFBoot2 with 1,000 replicates (Hoang et al. 2018). The trees were visualized with the ggtree package (version 2.2.1) (Yu et al. 2018). Multiple alignments and Newick tree files generated by phylogenetic analyses are available on FigShare (https://figshare.com/projects/EVEs-bias-virus-discovery/185650).

Statistical analysis

We assessed the amino-acid similarities between genomic EVE sequences and potential EVE sequences or exogenous viral sequences using the non-parametric two-sample Wilcoxon rank-sum test for paired data sets. We performed the same statistical test to evaluate disparities between the two groups for sequence lengths. Values of $P < 0.05$ were considered statistically significant. All statistical analyses were performed in the RStudio environment (v2021.09.1).

Results

Exploration of orthomyxovirus-associated EVEs in genomic datasets

To explore the impact of EVEs on virus discovery, we conducted a targeted case study examining orthomyxoviruses in Culicinae mosquitoes. Our initial step involved determining the presence of EVEs within thirteen publicly available genome assemblies containing the species *Culex pipiens* ($n = 1$), *Culex quinquefasciatus* ($n = 2$), *Aedes aegypti* ($n = 5$), and *Aedes albopictus* ($n = 5$), with some of them corresponding to strains and isolates widely used in cell culture (Supplementary Table 1). The screening identified orthomyxovirus-like sequences in ten of the thirteen genomes, all corresponding to the *Aedes* genus. We identified a total of 223 significant matches (e-values $<1 \times 10^{-4}$) to orthomyxoviruses in the genomes of *Aedes aegypti* and *Aedes albopictus* (Supplementary Table 3). No matches were found within *Culex* genomes. EVEs related to seventeen species were identified, all assigned to unclassified species within the genus *Quaranjavirus*. Most detected EVEs reveal numerous genomic sequences related primarily to the Guadeloupe mosquito quaranja-like virus, Usinis virus, and Aedes alboannulatus orthomyxo-like virus. The reported sequences have 26–91 per cent (average 51 per cent) amino acid identity with the present-day known exogenous orthomyxovirus proteins. Eighty-one per cent of EVEs were derived from the nucleoprotein (NP) gene ($n = 180$), while others were derived from the RdRP-associated segment PB1 ($n = 6$), the glycoprotein (GP) ($n = 9$), and hypothetical proteins 2 ($n = 23$) and 3 ($n = 5$) with currently unknown function (Fig. 2). Twelve EVEs related to the NP, GP, and hypothetical protein 3 encompass the entire length of the segment. Overall coverage lengths of EVEs compared to the corresponding reference sequence vary from 11.3 per cent to 100 per cent, representing partial and full.

In the predicted coding sequence, 122 sequences contained at least one frameshift mutation and 85 EVEs contained premature stop codons. However, seventy-three EVE hits (32.3 per cent of all genomic EVEs discovered) did not contain stop codons or frameshifts and can be considered intact full or partial ORFs (Supplementary Table 4). In addition, six hits were roughly the same size (>90 per cent amino-acid length) as the related viral segments.

In addition, the reference genomes and transcriptomes, later used for host read removal in transcriptomics dataset analyses, were examined for EVEs. *Culex* datasets showed no evidence of EVE sequences. However, in the *Aedes* datasets, thirty-three EVEs were

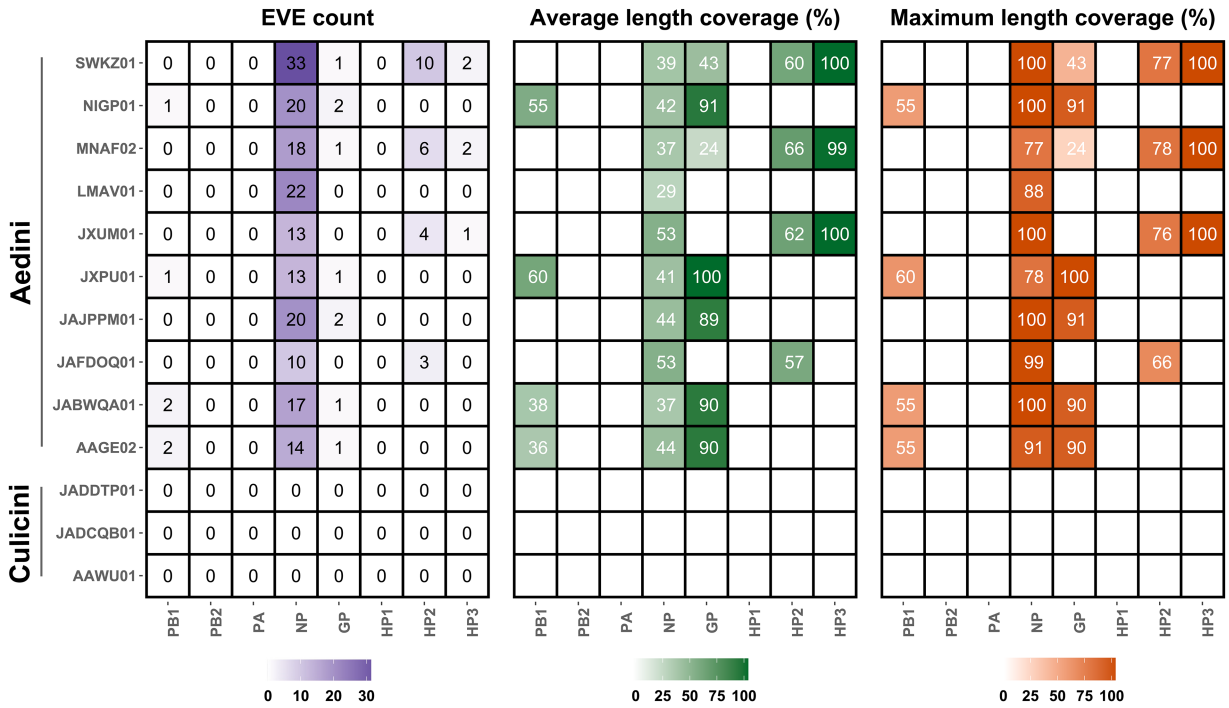

Figure 2. EVEs within genomic datasets exhibit bias toward nucleoprotein integration and showcase full ORF coverages. Tiles of heatmaps represent orthomyxovirus segments of 10 Aedini and 3 Culicini mosquito genome assemblies (top to bottom). Length coverages were calculated by dividing EVE hits' lengths (in amino acids) by the length (in amino acids) of the closest orthomyxovirus reference hit. Detailed descriptions of mosquito datasets and EVE length coverages can be found in Supplementary table 1 and 4, respectively. Abbreviations: PB1 = polymerase basic protein 1, PB2 = polymerase basic protein 2, PA = polymerase acidic protein, NP = nucleoprotein, GP = glycoprotein, HP1 = hypothetical protein 1, HP2 = hypothetical protein 2, HP3 = hypothetical protein 3.

identified in the genome, and the transcriptome contained four EVEs identical to the genomic EVEs (Supplementary table 5). These EVEs exhibited amino acid identities ranging from 24 per cent to 82 per cent to exogenous orthomyxovirus proteins. Among the identified EVEs, 84 per cent were derived from the NP gene ($n = 28$), while the remaining EVEs originated from the RdRP-associated segment PB1 ($n = 2$), hypothetical protein 2 ($n = 2$), and the GP ($n = 1$).

Screening for orthomyxoviruses in transcriptomic datasets

To understand how EVEs may affect virus detection in transcriptomic data, we sought to identify transcribed EVEs and external viral sequences associated with *Orthomyxoviridae* in publicly available RNA-Seq datasets. A total of 538 RNA-Seq libraries from 23 BioProjects were downloaded from the SRA NCBI database (Supplementary Table 6) (Leinonen, Sugawara, and Shumway 2011). Approximately half of the BioProjects ($n = 13$) have been previously used for virus detection (Batovska et al. 2019; Batson et al. 2021; Chandler et al. 2014; McBride et al. 2014; Shi et al. 2019; Ramos-Nino et al. 2020; Konstantinidis et al. 2022; Gil et al. 2023), with three known to have annotated orthomyxovirus accession entries (Supplementary table 7), but were screened nevertheless and considered controls for our virus discovery pipeline. Screened samples mainly represent wild-caught adults or larvae from individuals or pools from various geographic locations (Fig. 3A). Tribes included in the transcriptomic data are represented by 166 Aedini, 321 Culicini, 2 Culisetini, 5 Mansoniini, 3 Sabethini, 6 Toxorhynchitini, 1 Uranotaeniini, and 10 non-defined metatranscriptomic samples.

Orthomyxovirus-like contigs were present in ninety-four (~18 per cent of total datasets screened) samples, of which seventy-two samples have unannotated accessions for these viruses. These sequence hits were found in datasets belonging to the Culicinae tribes Aedini (42), Culicini (48), and non-defined metatranscriptomic samples (4) (Fig. 3B). *Culex* samples are primarily located in Europe. *Aedes* samples (mostly tropical species) are centered around the tropics and sub-tropics. As expected, previously annotated orthomyxovirus sequences in control datasets have been detected again by our pipeline.

The screening of sequencing libraries resulting from host read removal without integrated EVEs identified 2,494 contigs with orthomyxovirus-like hits (Supplementary tables 8 and 9). The majority of the contigs were taxonomically assigned to yet unclassified species of the genus *Quaranjavirus* detected primarily in mosquito hosts. Guadeloupe mosquito quaranja-like virus 1 and Wuhan Mosquito Virus 6 were the most represented, with over 600 hits each. We identified twenty-five samples (without controls) where we could detect all eight segments. However, additional orthomyxovirus contigs with distinct nucleotide similarities to the full-length segments were detected for segments in the majority of the samples (Fig. 4 and Supplementary Table 10). Seven Aedini samples comprised only sequences associated with the NP (Fig. 4), and thirty-four samples contained additional sequences per segment. Nineteen datasets associated with *Culex* mosquitoes contained only three segments or less. These sequences mostly showed low similarity to their best reference hit, incomplete fragment lengths, and often mapped to the same loci (data not shown). Besides NP, these were primarily associated with the RdRP catalytic subunit 1 (PB1), the GP, and the hypothetical proteins HP1 and HP3.

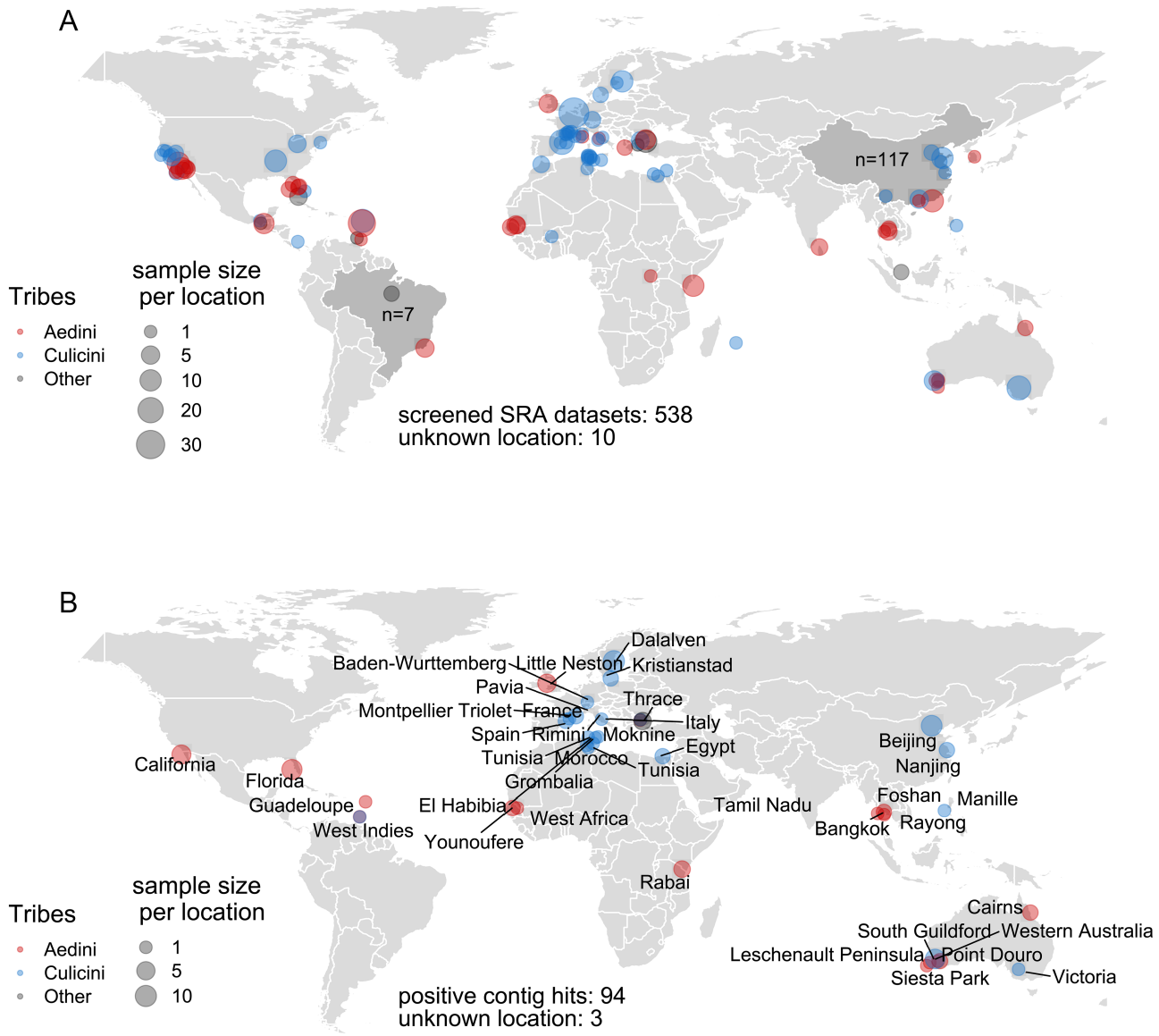

Figure 3. Positive samples of orthomyxoviruses demonstrate a global distribution pattern. Red, blue and grey dots indicate the sampling site, provided by the original study, of Aedes, Culex and other Culicinae transcriptomic datasets screened. Size represents sample count. Latitude and longitude positions were either taken from metadata or estimated according to the location name. a) Screened datasets used for virus discovery pipeline. Dark grey areas (Brazil, China) indicate additional screened samples with unknown coordinates. b) Samples with positive orthomyxovirus contig hits.

EVE detection in transcriptomic data

To explore the potential presence of EVE-like sequences within transcriptomic data analysis, we generated two custom databases consisting of either the already existing orthomyxovirus reference list or the translated amino-acid sequences of previously detected genomic EVEs. Since genomic EVEs were only found in *Aedes* assemblies, only contigs with orthomyxovirus hits derived from Aedini samples were used as a query for a diamond blastx search against the two databases. The highest bit score hits against each database were merged into a table and the difference between bit scores (Δ bit score) was calculated (Supplementary Table 11). The highest bit score values and Δ bit scores are plotted in Fig. 5A/B. High overall bit scores for non-EVE (exogenous orthomyxovirus references, maximum bit scores >500) sequences also display a higher Δ bit score (positive values) (Fig. 5A). This indicates that putative exogenous sequences have hits with longer length matches to non-EVE references (paired samples Wilcoxon signed-rank test, $P < 2.2e\text{-}16$), shorter length matches to EVE sequences, and significantly lower amino-acid similarities to EVEs (paired samples Wilcoxon signed-rank test, $P < 2.2e\text{-}16$). These long-length matches are, with a few exceptions, missing for hits with a better bit score for genomic EVE sequences.

Since most of the screened datasets were not generated for virus detection, the quality of viral sequence reads is generally low, which leads to shorter read assemblies. Consequently, maximum bit scores are low and show little difference in amino-acid identity when aligned to EVEs or non-EVEs ($-100 < \Delta$ bit score < 100). As different library preparation techniques were used for different BioProjects, bit score data points were analyzed by BioProject (Supplementary Fig. 1), but no noticeable relationship between bit score distributions and virus- or non-virus-focused studies could be seen. Figure 5 also depicts detected frameshifts or nonsense

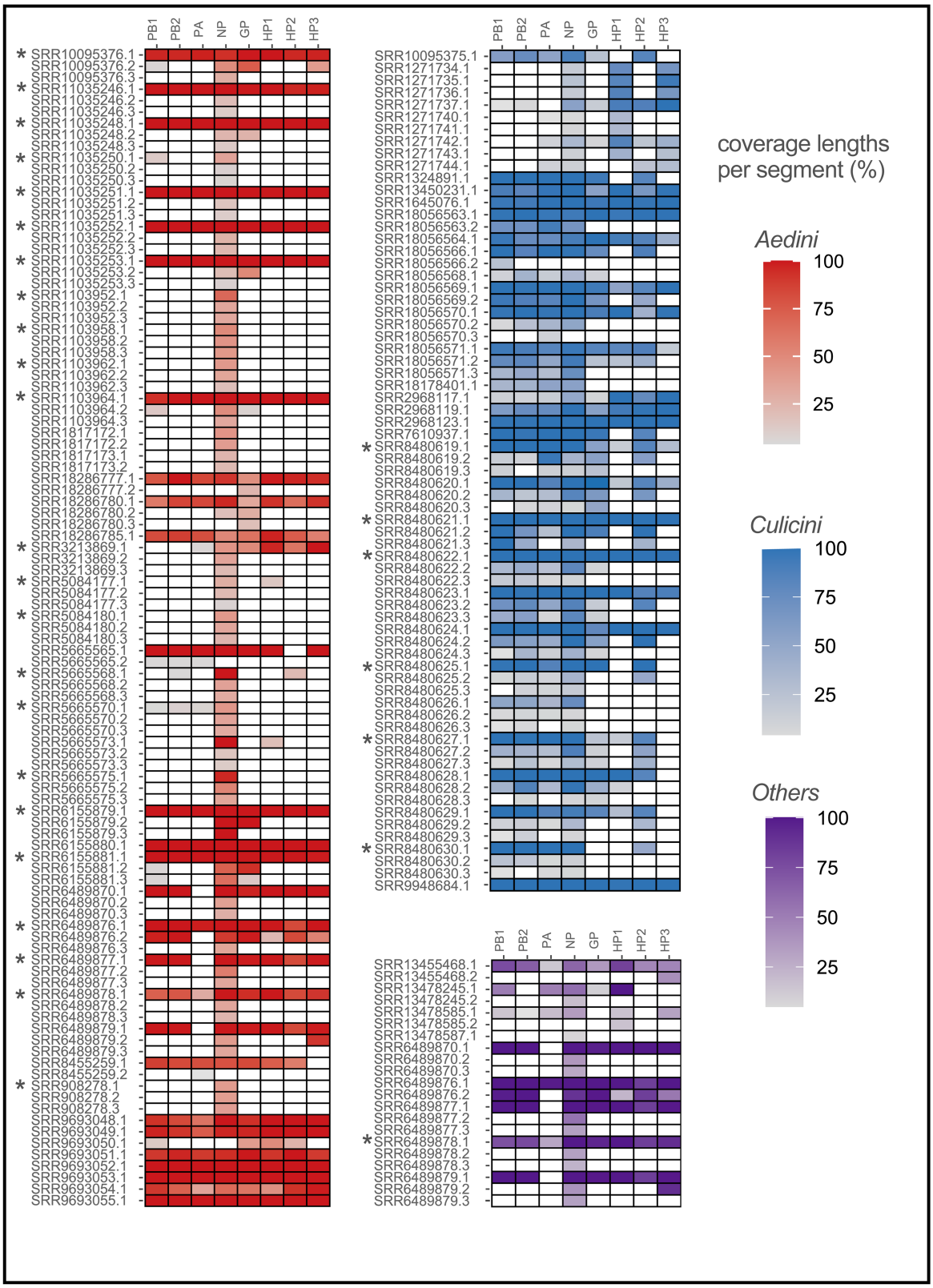

Figure 4. Orthomyxovirus detection in SRA datasets reveals an abundance of isolated segments. Here, we assess the completeness of our contigs for each viral segment. Contigs are aligned to each segment's best matching accession hits to check for full or partial segment lengths. Additional contigs with high sequence variety to an already aligned sequence per segment were separated and treated as additional segments. For each segment, length coverages were calculated. Aedini, Culicini and other Culicinae samples are colored in a red, blue, or violet gradient according to their coverage percentages, respectively. Segments per sample were ordered according to their highest percentages and are represented as tiles. Appended numbers to accession numbers (.2, .3) represent additional sequences per segment per sample, with up to three additional sequences shown in this figure. Samples with more than three sequences per segment are annotated with an asterisk and complete coverage lengths for sequences can be found in Supplementary Table 10. Segments with no associated contigs are depicted in white.

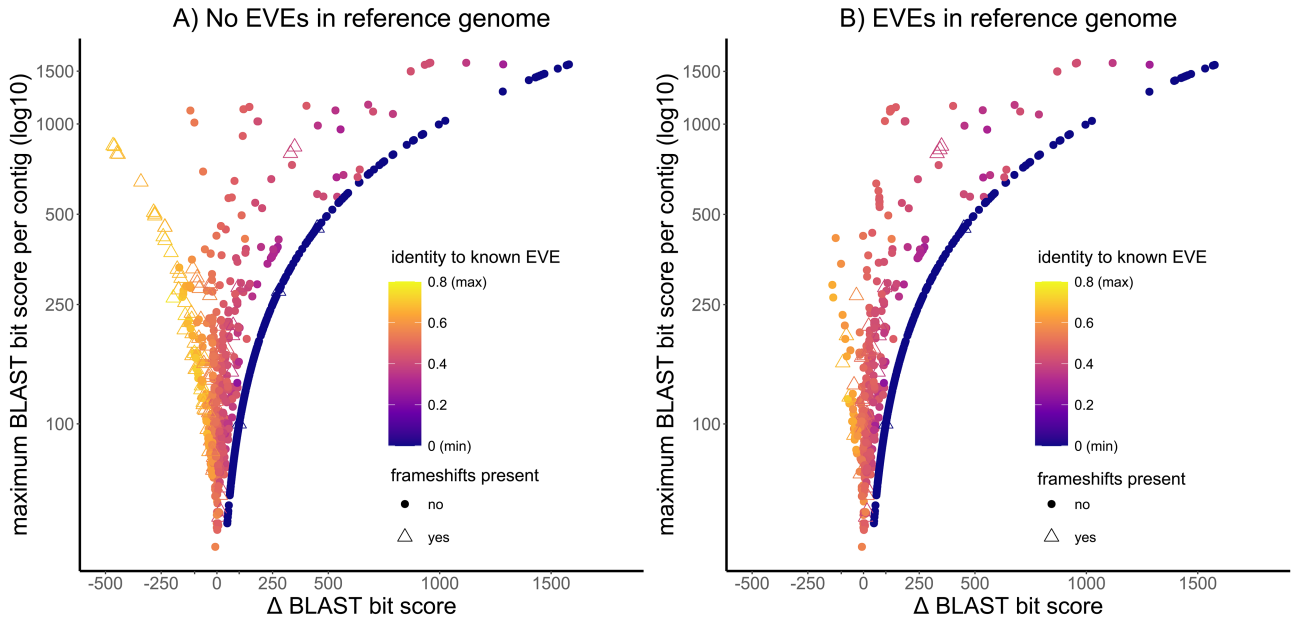

Figure 5. Bit score comparisons of SRR contig hits against EVE or exogenous viral sequences show reduced hits for pipeline with host read removal step, including EVE sequences. Contigs were blasted against two custom databases: the existing orthomyxovirus reference list or the translated amino-acid sequences of previously detected genomic EVEs. Data points are colored according to amino-acid sequence similarity to EVE references. Contigs with frameshifts or nonsense mutations (stop codons) are indicated with triangular shapes. Contigs after host read removal without integrated EVEs in the reference genome are shown in panel A. Contigs after host read removal with present EVEs in the reference genome are shown in panel B. Details of the bit score calculations and frameshift corrections can be found in Supplementary Table 11.

mutations in *Aedes*-associated contigs. While most of such mutations occur at negative Δ bit scores (more EVE-related), some contigs with higher similarity to exogenous reference sequences also exhibit frameshifts. As bit score calculations are dependent on identified genomic EVEs, false positives for exogenous-like sequences are to be expected.

Influence of EVEs on host read removal

Upon host read removal from reference sequences with EVEs, 213 viral contigs were lost compared to contigs generated from reads passing host read filtering without EVEs. As shown in Fig. 5B, most of the lost contigs ($n = 107$) have higher blast bit scores for genomic EVEs (Δ bit score lower/equal −10) than exogenous orthomyxovirus sequences. However, forty-five contigs associated with the exogenous virus (Δ bit score greater than −10 + no frameshift/stop codons) were also lost. Additionally, not all EVE-associated sequences were lost due to host read removal, as contigs with frameshift/nonsense mutations can still be seen within the plot (Fig. 5B).

To avoid chimeric assemblies between putative EVE and non-EVE sequences during the post-processing analysis, a Δ bit score cut-off was used. Contigs with a Δ bit score lower than −10 and frameshift mutations ($n = 179$) were set aside as potential EVE sequences and excluded from manual contig extensions. Out of these, eighty-four contained frameshift mutations, and twelve contained nonsense mutations (stop codons) (Supplementary table 11). Following contig extension and post-processing steps detailed in the 'Material and methods' section, 452 previously unreported non-EVE orthomyxoviral sequences were generated.

Influence of EVEs on read coverage and SNP analysis

For a more profound understanding of contig loss during host read removal, we closely inspect this process at the read level, conducting read coverage analysis across all four workflows. After mapping on the host genome containing EVEs, a reduction of read abundance can be seen in twenty-two NP segments (68 per cent of *Aedes* samples). Upon comparison of host read removal with the different transcriptomic references, no changes in sequencing depth were observed. We compared nucleotide similarities between genomic EVEs and our transcriptomic datasets to investigate further the potential bias of EVE presence in host read removal approaches. Regions with reduced read depth were isolated and aligned against EVE sequences found in the reference genome. Best-aligned hits can be seen in Supplementary Table 12. Nucleotide similarities ranged between 74 per cent and 84.55 per cent. To elucidate the impact of read reduction on the overall loss of sequence information, coverage comparisons between the four generated sequences from each workflow were done and can be found in Supplementary Fig. 2. Three sequences exhibited regions within the segments that fell below the sequencing depth threshold after genomic host read removal with EVE presence. Lastly, single nucleotide polymorphisms (SNPs) were explored among the four different pipelines. While SNPs could be detected in twenty datasets, most can be attributed to equally abundant variants chosen stochastically. However, four SNPs generated within regions of reduced read depth show a switch between major/minor variants and are therefore attributed to differences in host mapping (SNP distributions in Supplementary Fig. 2).

Phylogeny and implications on taxonomy

To illustrate the evolutionary relationships between exogenous orthomyxoviruses and EVEs, we aligned putative EVE protein sequences with viral protein sequences obtained after removing host reads containing EVEs. We then constructed maximum likelihood phylogenetic trees for each genetic segment (Supplementary Fig. 3–9, Fig. 6). The NP was chosen as representative, as most SRA samples and EVEs correspond to this segment (Fig. 6). The

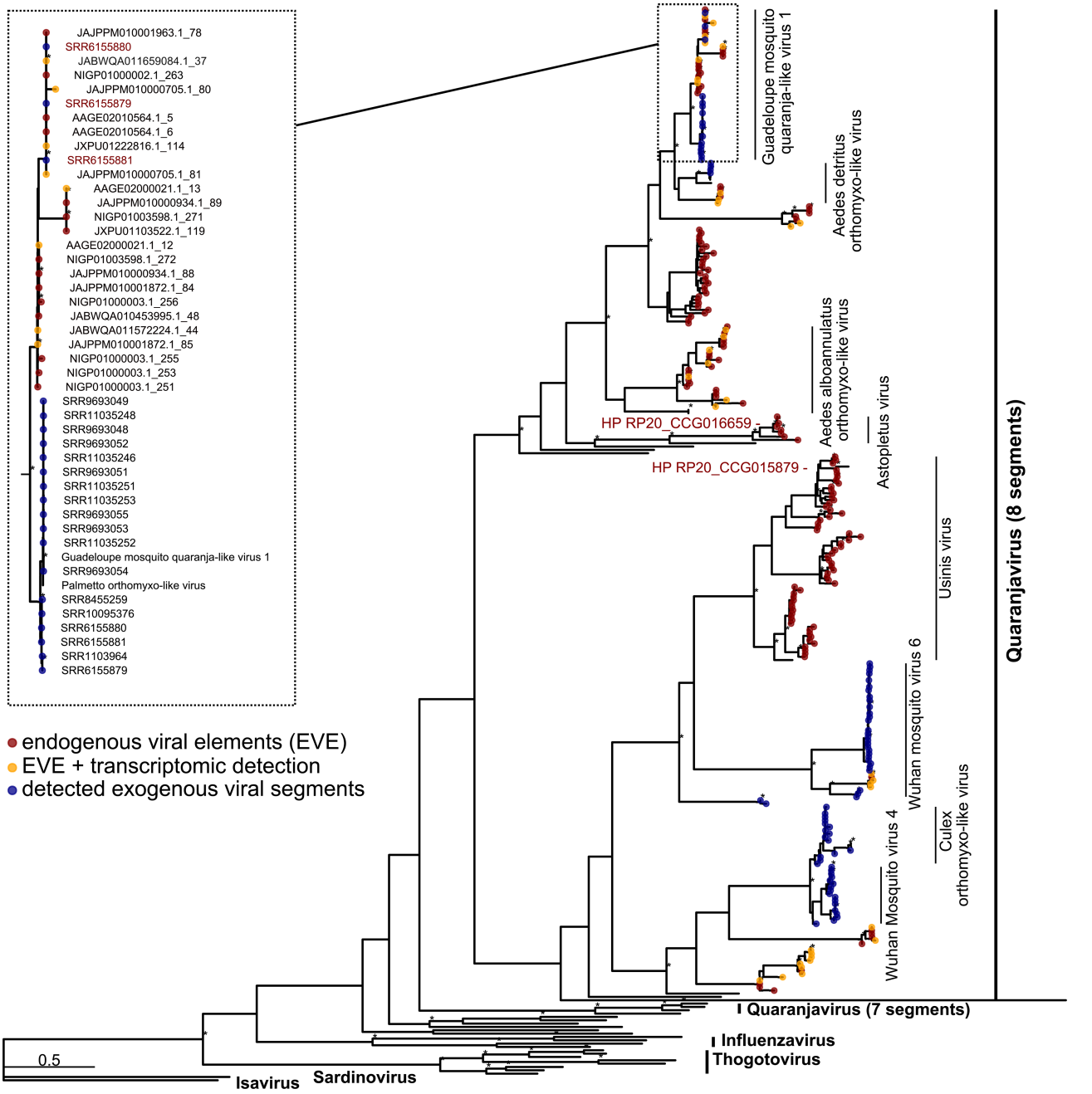

Figure 6. Phylogenetic tree depicting relationships between NP sequences of orthomyxoviruses. Sequence alignments of genomic EVEs, transcriptomic orthomyxovirus sequences, and references were generated with mafft v7.490. The maximum likelihood tree was constructed with model VT+I+G4 in IQ-TREE 1.6.12 (+ ultrafast bootstrap (1000 replicates)). Significant ultrafast bootstrap values >95% are depicted with an asterisk. Protein sequences detected in this study are marked as colored circles at the tips: red - genomic EVEs, orange - genomic EVEs + detection of transcriptomic contigs with sequence similarity >95% and a Δbit score <-10, dark blue - putative exogenous sequences. Species and genus (bold) characteristics per clade are highlighted with black vertical lines to the right of the tree. Transcribed EVE sequences with fully intact ORF are depicted in bold red. Two hypothetical proteins from Aedes albopictus (HP RP20) are pointed out in red. Multiple sequence alignment corresponding to NP phylogenetic tree can be found on FigShare https://figshare.com/projects/EVEs-bias-virus-discovery/185650.

final sequences produced in this study clustered in two distinct clades, forming a sister lineage to the classified genus *Quaranjavirus*. The first clade (top) consists of primarily *Aedes*-associated orthomyxoviruses except for Astopletus virus, a recently identified Quaranjavirus detected in Culicini. Eighteen transcriptomic samples clustered with Guadeloupe mosquito quaranja-like virus 1, detected initially in Guadeloupe and California (Batson et al. 2021; Shi et al. 2019). Six samples clustered to Aedes detritus orthomyxo-like virus and one to Aedes alboannulatus orthomyxo-like virus. For the second clade (bottom), nineteen datasets include sequences identified as Wuhan Mosquito virus 6 and Culex pipiens orthomyxo-like virus. While these two references have been deposited with different names and are derived from different continents, they share a >99 per cent amino-acid similarity. Four datasets from the same BioProject clustered on a separate branch to the known Wuhan mosquito viruses 6. Sixteen sequences were

designated as Wuhan Mosquito virus 4-like and sixteen sequences clustered with Culex orthomyxo-like virus.

Genomic EVEs cluster within distinct sister clades to their exogenous viral counterparts and are closely related to each other. The twenty-nine EVEs form three subgroups related foremost to the Guadeloupe mosquito quaranja-like virus 1. Six integrations show the highest similarity to the Aedes detritus orthomyxo-like virus, while twenty-six EVEs form a sister clade to the common ancestor of the two previously mentioned viral species. Twenty-five EVEs cluster with Aedes alboannulatus orthomyxo-like virus, and seven EVEs cluster with the recently described Astopletus virus. With fifty-seven EVEs detected for Usinis virus, these sequences formed the most extensive clade of EVEs within the tree. Surprisingly, no new potential exogenous samples and putative transcriptomic EVEs could be found for the Usinis virus. Five and nineteen EVEs were closely related to Wuhan Mosquito virus 6 and 4, respectively. Astopletus virus, Wuhan Mosquito virus 4 and 6 have only been detected in *Culex* mosquitoes previously. However, all detected genomic EVEs were detected in *Aedes* genomic data, which suggests that they are or were able to also infect *Aedes* mosquitoes persistently.

In addition to genomic EVEs, putative transcribed endogenous viral contigs, previously excluded for further downstream analysis, with a high sequence similarity to genomic EVEs, were also visualized in the tree (depicted in orange). The excluded contigs (n = 209), had the highest similarities to forty-two EVEs, all derived from *Aedes aegypti* host genomes, and can be found evenly distributed among the genomic EVE clades (except Usinis virus). To illustrate a lack of putative EVE filtering, three consensus sequences generated of putative EVE contigs (Δ bit score <−10) were included for subsequent sequence correction, coverage analysis, and tree construction. All consensus sequences depicted full-length contigs for an additional NP segment to an already complete orthomyxovirus genome (all eight segments present) (Supplementary Figure 10), but also included extended sequences with blast hits to uncharacterized insect proteins, validating them as transcribed EVEs. As seen in Fig. 6, the sequences, named after their SRA datasets SRR6155879, SRR6155880, and SRR6155881 (bold red), cluster within the genomic EVEs related to Guadeloupe mosquito quaranja-like virus 1. Upon closer inspection of their amino acid sequence, all sequences share 100 per cent similarity to a detected genomic EVE with >98 per cent segment coverage and a mean depth of 10, 22.6, and 36.2. All three sequences represent a fully intact ORF, share 100 per cent identity, and, although derived from the same BioProject, differ in sampling locations (Cairns, Australia and Bangkok, Thailand), indicating a highly conserved integration event across geographically diverse samples.

While screening for genomic and transcriptomic integration events, various orthomyxoviral sequences had the highest or second-highest blast hits to hypothetical proteins annotated as part of the *Aedes albopictus* genome. This is true for EVEs related to Astopletus and Usinis viruses and both recently described Quaranja-like viruses with eight segments (Batson et al. 2021). These proteins (NCBI accession numbers: KXJ73423.1 and KXJ73043.1) are likely orthomyxovirus EVEs, falsely annotated as part of the host genome (Fig. 6 and Supplementary Figure 11).

Discussion

EVEs and virus discovery

Viral metagenomic analysis has become an important approach in uncovering novel viral genomes in various samples, giving an ever-increasing insight into the diversity of the virosphere. With the rapid pace of improved sequencing techniques and assembly tools, viral sequence detection in samples has become sufficient evidence for admission as *bona fide* viruses, no longer relying on the verification through phenotypic properties.

While metagenomics and metatranscriptomics are undoubtedly the present and future of virus discovery, appropriate checks for data accuracy and integrity are essential to infer the actual existence of a virus with solely sequence data. Unfortunately, these steps are not standardized among virus discovery pipelines, leading to discrepancies regarding coverage and sequencing depth thresholds, ensuring the validity of viral sequences. In addition, more and more studies focus on re-analyzing samples initially intended for non-virus-related applications. Before and during library preparation, suboptimal processing often correlates with poor viral RNA quality, leading to lower read abundance and shorter assembled contigs. Complete genome assemblies are difficult, and scaffold assemblies and contig extensions are often necessary to reach full genome length. These assembled sequences risk being derived from multiple virus populations, leading to artificially generated chimeric genomes, which can occur between closely related exogenous virus populations and with integrated viral sequences in the host if their sequences are sufficiently similar.

Non-retroviral EVEs in the sequenced host genome are often ignored in virus discovery pipelines, and bioinformatic steps undertaking EVE removal are rare. RNA-Seq libraries especially forgo this additional downstream cleaning step, as DNA removal steps and mRNA enrichment techniques are often part of cDNA library preparations. However, some EVEs are transcribed and present among the sequenced transcribed RNAs (Katzourakis, Gifford, and Malik 2010). Our study shows that an important number of RNA sequences identified from RNA-Seq libraries have high amino acid similarities with EVEs detected from corresponding genomic host datasets. While most of these sequences only consist of partial segments or contain frameshift mutations, some of them can span whole segments and display intact ORFs and sufficient abundance levels. Without initial EVE removal steps, these sequences can be easily misinterpreted as belonging to exogenous orthomyxoviruses, as seen in this study. Samples with low-quality viral contigs are challenging to classify as either more EVE- or exogenous-like, especially considering that some genomic EVEs were identical to their respective exogenous counterparts, with intact full-length ORFS, and/or expressed. Without a host genome and preliminary EVE characterization, and in a context of low viral reads abundance, which is common in metagenomic studies, this can lead to the misidentification of an EVE as an exogenous virus, especially when sequence assemblies contain incomplete genomes of novel viruses. We thus suggest that partial genomes or a limited number of viral segments should be viewed cautiously to avoid a potential misidentification in sequence origin.

When a host genome is available, reads are typically mapped to a host genome or transcriptome reference to alleviate the time-consuming assembly tasks. While this limits the presence of EVEs in the assembled viral contigs, it can also lead to loss of information as virus-associated reads can match genomic EVEs, as we have shown in this study. This can be especially detrimental for sup-optimal samples without virus enrichment, as read depths are usually low and could get lost during quality cut-offs. Despite not being a common event, integrated EVEs in the reference genome might also alter the final consensus sequences of the present viral population. This further complicates the already challenging task of distinguishing mixed infections or viral population clouds as potential

variations at single nucleotide positions could be observed in our data.

Orthomyxoviridae-derived EVEs in mosquitoes

Despite the limited sampling size of non-retroviral EVEs, most seem to arise from negative-strand RNA viruses (Holmes 2011; Aiewsakun and Katzourakis 2015; Blair, Olson, and Bonizzoni 2020). The exact mechanism of this favoritism of ssRNA(-) viruses is unclear, but one contributing factor could be the replication of certain viral families, such as *Bornaviridae* and *Orthomyxoviridae*, within the nucleus (Horie et al. 2010). This replication location might increase the likelihood of reverse transcription and subsequent genomic integration. It has also been suggested that the tendency of ssRNA(-) viruses to generate a large number of short mRNAs, in contrast to ssRNA(+) viruses, which typically produce fewer long mRNAs encoding a single polyprotein, might favor their endogenization (Holmes 2011). This increased production of RNA molecules increases the likelihood of undergoing reverse transcription and subsequent integration into the host's genetic material.

Surprisingly, the identification of orthomyxovirus-derived EVEs in insects has been relatively limited (Katzourakis, Gifford, and Malik 2010; Li et al. 2015; Russo et al. 2019; Palatini, et al., 2022). Notably, only one RdRp EVE has been described in *Aedes aegypti* so far (Li et al. 2015), but we are unable to compare this sequence to the ones reported here, as no sequences or contig locations have been provided. This relative scarcity of information can be attributed to the historically biased exploration of orthomyxoviruses towards pathogenic genera, predominantly *Influenzavirus* and *Thogotovirus*, which have never been identified as exogenous viruses in mosquitoes. Recent studies, however, have highlighted the widespread presence of orthomyxoviruses in arthropods (Batson et al. 2021; Shi et al. 2016). Our investigation incorporated a diverse array of recently described insect-specific orthomyxovirus species as reference input. Interestingly, most EVEs identified in our study are associated with orthomyxoviruses characterized in the last five years (Shi et al. 2019; Batson et al. 2021).

Similarly to EVE-derived from other viral families, detected orthomyxovirus-derived EVEs were found in *Aedes* mosquito genomes. Despite the high prevalence of orthomyxovirus sequences in transcriptomic *Culex* datasets, no EVEs were detected in *Culex* genomes. The reason for such low levels of viral integration is not known yet but is speculated to correlate to the presence of transposable elements (TEs) (Whitfield et al. 2017). However, while *C. quinquefasciatus* genome is known to comprise only 29 per cent of TEs compared to *Aedes aegypti* (42–47 per cent), the *Anopheles gambiae* genome has even fewer TEs (11–16 per cent), but shows a higher number of EVEs (Holt et al. 2002; Nene et al. 2007; Arensburger et al. 2010; Blair, Olson, and Bonizzoni 2020).

As seen in other EVE studies, we observed preferential integration of viral genes or segments: sequences of EVEs encode for the viral RNA-dependent-RNA-polymerase subunit PB1, the nucleocapsid (NP), a GP and two hypothetical proteins. Considering that over 80 per cent of the observed EVEs correspond to NP genes, a selective pressure to endogenize and/or keep this viral genomic region might be hypothesized. NP, the most abundant protein in infected cells (Kummer et al. 2014), has an essential role in viral RNA replication and transcription, organization of RNA packing, and nuclear trafficking (Herz et al. 1981; Martin and Helenius 1991; Eisfeld, Neumann, and Kawaoka 2015). NP endogenization might thus result from relative mRNA abundance in the nucleus. Preferential integration of NP, but also polymerase genes, was also seen in other negative-sense RNA viruses (Katzourakis et al. 2007; Katzourakis, Gifford, and Malik 2010; Holmes 2011; Gilbert and Belliardo 2022). This might be because, in most *Mononegavirales* species, the 3′ NP gene is notably abundant due to a directional stepwise transcription process starting from the 3′ end (Whelan, Barr, and Wertz 2004; Holmes 2011). However, opposite results have also been observed (Russo et al. 2019) and since genome organization and replication mechanisms differ between virus families, the drivers of endogenization need to be explored further.

Limitations in EVE detection

The quality of the host genomes considered in this study can play a pivotal role in the reliability and interpretability of the discovered EVE sequences. High-quality genome assemblies can produce chromosome-length scaffolds with long contigs and a minimal number of gaps. In contrast, lower-quality genomes often consist of numerous unanchored scaffolds with short contig lengths and significant gaps, which often harbor errors, misassemblies, and incomplete sequences, increasing the risk of false positives and negatives in EVE predictions. Further issues might arise with genomes derived from reference-guided assemblies: despite a simplified assembly process, they may exhibit biases towards the reference, including potentially misassembled EVEs, contamination by exogenous viruses, and missing highly repetitive or complex genomic regions. Additionally, different versions of the same EVEs, known as haplotypic nrEVE variants, can be present in the genomic dataset (Palatini et al. 2020), possibly due to pooled sample input. The extent to which input material and low-quality genome assemblies may introduce bias in EVE detection is not currently understood and illustrates the need for future research to explore the precise influence of genome quality on EVE detection. Finally, the similarity-based screening strategy usually employed for EVE detection, such as in this study, is likely to underestimate the true diversity of EVEs. Indeed, our screening process relies on our current and limited knowledge of viral diversity within *Orthomyxoviridae*, which can lead to the inadvertent omission of certain very divergent EVEs from the analysis. Furthermore, identifying ancient EVEs may be particularly challenging due to genetic divergence tied to changes in evolutionary pressures in both host and virus lineages, which can hide their presence.

Conclusions

Misidentification of sequences is expected to occur, considering the continuing exploration of viral sequencing data with metagenomics and metatranscriptomics. As seen in this study, annotation errors have already been detected in the genome of *Aedes albopictus*, with some hypothetical proteins being, in reality, integrated viral elements related to two recently discovered quaranjaviruses. In addition, misclassifying EVEs with exogenous viral sequences could contaminate databases and adversely affect downstream, e.g. phylogenetics, analyses.

However, accurate discrimination between EVEs and exogenous viral sequences remains complex, necessitating a comprehensive and multifaceted approach. As we have demonstrated in this study, read mapping onto host genomes, when available, can mitigate biases, albeit at the expense of sensitivity. When possible, a more effective and recommended strategy would be to exploit the genome assembly data of the host organism to screen for genomic EVEs. This can then facilitate the identification

and subsequent exclusion of highly similar sequences from transcriptomic datasets. This method proved effective not only in our study but also in the recently described case of the potato aphid *Macrosiphum euphorbiae*, where a previously described exogenous *Ambidensovirus* was confirmed to be, in reality, an actively transcribed EVE (Rozo-Lopez et al. 2023).

In cases where host genomes or genomic sequence data are unavailable, several alternative strategies can aid in the differentiation between EVEs and exogenous viruses. One would involve the identification of disrupted ORFs or nonsense mutations, characteristics linked to EVEs with a history of long-term integration, although it may not apply to recently integrated or functional EVEs. An additional discriminatory approach relies on quantifying transcript abundance, contrasting putative viral sequence expression levels to host housekeeping genes. However, this method will not reliably detect actively transcribed EVEs with specific functions, such as those involved in viral immunity, which might show increased expression during a viral infection. The inspection of presumed viral reads for host DNA presence can also prove informative, given the colocalization of EVEs with host genetic material, but might be biased by technical errors (Peccoud et al. 2018). Furthermore, comparing the phylogenetic affiliations between putative viral sequences and established EVEs can offer additional insights, as EVEs typically form distinct clusters, separate from exogenous viruses. Integration of small RNA sequencing data can further aid EVE identification, as many species exhibit an enrichment of EVEs within P-element-induced wimpy testis (PIWI)-interacting RNA clusters (Ter Horst et al. 2019). While each technique has limitations, we recommend employing a combined strategy integrating multiple approaches.

While distinct features differentiating EVEs from exogenous viruses exist, no standalone tools combining these criteria are currently available for EVE detection. Computational aids like CheckV (Nayfach et al. 2021) can partially filter EVE sequences by discerning host gene contamination within virus-like sequences (Mifsud et al. 2022; Costa et al. 2023); however, applications are constrained by stringent sequence quality prerequisites. Importantly, despite the multiple strategies for precise EVE–exogenous distinction, complete classification into either category remains elusive solely through sequencing data. Nevertheless, the differentiation of EVEs from exogenous sequences is crucial, offering insights into the historical interplay between viruses and their hosts, thus illuminating the impact of EVEs on host physiology and immunity.

While there is no doubt that today's virus discovery technologies heavily rely on metagenomic sequencing, it is essential to consider possible limitations due to the presence of EVEs in transcriptomic and meta-transcriptomic datasets. Datasets with a low viral abundance are especially susceptible to bias, primarily due to the challenge of distinguishing between exogenous and endogenous viral sequences. We argue that additional caution should be taken upon detecting novel virus sequences and that a framework for EVE detection should be a standard step within virus discovery pipelines and would provide great validation of exogenous viral sequences derived from metagenomic sequencing.

Data availability

Sequence files, supplementary data, tools, versions, and scripts used for analysis in this study have been deposited in Figshare (https://figshare.com/projects/EVEs-bias-virus-discovery/185650). No raw sequencing data was generated for this study. Viral genomes and their NCBI GenBank under accession numbers can also be accessed in the FigShare repository.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Conflict of interest: None declared.

Author contributions

NB and SL conceived the study. NB performed formal analyses and contributed to visualization. NB, TH, and SL contributed to validation. CM, AE, and SG provided sequence input. SL contributed to supervision, project administration, and funding acquisition. NB, TH, and SL wrote the original draft of the manuscript. All authors read and approved the manuscript.

References

Ackermann, H. W., and DuBow, M. S. (1987) Viruses of rokaryotes: *General Properties of Bacteriophages*, 1. pp. 49–85, Boca Raton, Florida: CRC Press, Viruses of Prokaryotes.

Aiewsakun, P., and Katzourakis, A. (2015) 'Endogenous Viruses: Connecting Recent and Ancient Viral Evolution', *Virology*, 479–480: 26–37.

Allison, A. B. et al. (2015) 'Cyclic Avian Mass Mortality in the Northeastern United States Is Associated with a Novel Orthomyxovirus', *Journal of Virology*, 89: 1389–403.

Amos, B. et al. (2022) 'VEuPathDB: The Eukaryotic Pathogen, Vector and Host Bioinformatics Resource Center', *Nucleic Acids Research*, 50: D898–911.

Arensburger, P. et al. (2010) 'Sequencing of Culex Quinquefasciatus Establishes a Platform for Mosquito Comparative Genomics', *Science (New York, N.Y.)*, 330: 86–8.

Batovska, J. et al. (2019) 'Sensitivity and Specificity of Metatranscriptomics as an Arbovirus Surveillance Tool', *Scientific Reports*, 9: 19398.

Batson, J. et al. (2021) 'Single Mosquito Metatranscriptomics Identifies Vectors, Emerging Pathogens and Reservoirs in One Assay', *eLife*, 10: e68353.

Bejarano, E. et al. (1996) 'Integration of Multiple Repeats of Geminiviral DNA into the Nuclear Genome of Tobacco during Evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 93: 759–64.

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010) 'Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes', *PLoS Pathogens*, 6: e1001030.

Benson, D. A. et al. (2013) 'GenBank', *Nucleic Acids Research*, 41: D36–42.

Benveniste, R. E., and Todaro, G. J. (1974) 'Evolution of C-type Viral Genes: Inheritance of Exogenously Acquired Viral Genes', *Nature*, 252: 456–9.

Berns, K. I., and Linden, R. M. (1995) 'The Cryptic Life Style of Adenoassociated Virus', *BioEssays*, 17: 237–45.

Blair, C. D., Olson, K. E., and Bonizzoni, M. (2020) 'The Widespread Occurrence and Potential Biological Roles of Endogenous Viral Elements in Insect Genomes', *Current Issues in Molecular Biology*, 34: 13–30.

Chandler, J. A. et al. (2014) 'Metagenomic Shotgun Sequencing of a Bunyavirus in Wild-caught Aedes Aegypti from Thailand Informs the Evolutionary and Genomic History of the Phleboviruses', *Virology*, 464–465: 312–9.

Costa, V. A. et al. (2023) 'Limited Cross-species Virus Transmission in a Spatially Restricted Coral Reef Fish Community', *Virus Evolution*, 9: vead011.

Crochu, S. et al. (2004) 'Sequences of Flavivirus-related RNA Viruses Persist in DNA Form Integrated in the Genome of Aedes Spp. Mosquitoes', *Journal of General Virology*, 85: 1971–80.

Edgar, R. C. et al. (2022) 'Petabase-scale Sequence Alignment Catalyses Viral Discovery', *Nature*, 602: 142–7.

Eisfeld, A. J., Neumann, G., and Kawaoka, Y. (2015) 'At the Centre: Influenza A Virus Ribonucleoproteins', *Nature Reviews, Microbiology*, 13: 28–41.

García-López, R., Vázquez-Castellanos, J. F., and Moya, A. (2015) 'Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations', *Frontiers in Bioengineering and Biotechnology*, 3: 141.

Geisler, C., and Jarvis, D. L. (2016) 'Rhabdovirus-like Endogenous Viral Elements in the Genome of Spodoptera Frugiperda Insect Cells are Actively Transcribed: Implications for Adventitious Virus Detection', *Biologicals: Journal of the International Association of Biological Standardization*, 44: 219–25.

Geuking, M. et al. (2009) 'Recombination of Retrotransposon and Exogenous RNA Virus Results in Nonretroviral cDNA Integration', *Science (New York, N.Y.)*, 323: 393–6.

Gil, P. et al. (2023) 'Spatial Scale Influences the Distribution of Viral Diversity in the Eukaryotic Virome of the Mosquito Culex Pipiens', *Virus Evolution*, 9: vead054.

Gilbert, C., and Belliardo, C. (2022) 'The Diversity of Endogenous Viral Elements in Insects', *Current Opinion in Insect Science*, 49: 48–55.

Herniou, E. et al. (1998) 'Retroviral Diversity and Distribution in Vertebrates', *Journal of Virology*, 72: 5955–66.

Herz, C. et al. (1981) 'Influenza Virus, an RNA Virus, Synthesizes Its Messenger RNA in the Nucleus of Infected Cells', *Cell*, 26: 391–400.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Holmes, E. C. (2011) 'The Evolution of Endogenous Viral Elements', *Cell Host & Microbe*, 10: 368–77.

Holt, R. A. et al. (2002) 'The Genome Sequence of the Malaria Mosquito *Anopheles Gambiae*', *Science*, 298: 129–49.

Horie, M. et al. (2010) 'Endogenous Non-retroviral RNA Virus Elements in Mammalian Genomes', *Nature*, 463: 84–7.

Jaenisch, R. (1976) 'Germ Line Integration and Mendelian Transmission of the Exogenous Moloney Leukemia Virus', *Proceedings of the National Academy of Sciences of the United States of America*, 73: 1260–4.

Johansen, J. et al. (2022) 'Genome Binning of Viral Entities from Bulk Metagenomics Data', *Nature Communications*, 13.

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Katzourakis, A. et al. (2007) 'Discovery and Analysis of the First Endogenous Lentivirus', *Proceedings of the National Academy of Sciences*, 104: 6261–5.

Katzourakis, A., Gifford, R. J., and Malik, H. S. (2010) 'Endogenous Viral Elements in Animal Genomes', *PLoS Genetics*, 6: e1001191.

Katzourakis, A., Rambaut, A., and Pybus, O. (2005) 'The Evolutionary Dynamics of Endogenous Retroviruses', *Trends in Microbiology*, 13: 463–8.

Konstantinidis, K. et al. (2022) 'Defining Virus-carrier Networks that Shape the Composition of the Mosquito Core Virome of a Local Ecosystem', *Virus Evolution*, 8: veac036.

Krishnamurthy, S. R., and Wang, D. (2017) 'Origins and Challenges of Viral Dark Matter', *Virus Research*, 239: 136–42.

Kummer, S. et al. (2014) 'Alteration of Protein Levels during Influenza Virus H1N1 Infection in Host Cells: A Proteomic Survey of Host and Virus Reveals Differential Dynamics', *PLoS ONE*, 9: e94257.

Leinonen, R., Sugawara, H., and Shumway, M., International Nucleotide Sequence Database Collaboration. (2011) 'The Sequence Read Archive', *Nucleic Acids Research*, 39: D19–21.

Lequime S and Lambrechts L (2017) Discovery of flavivirus-derived endogenous viral elements in Anopheles mosquito genomes supports the existence of Anopheles-associated insect-specific flaviviruses *Virus Evol* 3 vew035

Li, C.-X. et al. (2015) 'Unprecedented Genomic Diversity of RNA Viruses in Arthropods Reveals the Ancestry of Negative-sense RNA Viruses', *eLife*, 4: e05378.

Li Y *et al* (2022) Endogenous Viral Elements in Shrew Genomes Provide Insights into Pestivirus Ancient History *Molecular Biology and Evolution* 39 msac190

Liu, H. et al. (2011) 'Widespread Endogenization of Densoviruses and Parvoviruses in Animal and Human Genomes', *Journal of Virology*, 85: 9863–76.

Liu, W. et al. (2012) 'The First Full-Length Endogenous Hepadnaviruses: Identification and Analysis', *Journal of Virology*, 86: 9510–3.

Marchler-Bauer, A. et al. (2015) 'CDD: NCBI's Conserved Domain Database', *Nucleic Acids Research*, 43: D222–226.

Martin, K., and Helenius, A. (1991) 'Transport of Incoming Influenza Virus Nucleocapsids into the Nucleus', *Journal of Virology*, 65: 232–44.

Matthews, B. J. et al. (2018) 'Improved Reference Genome of Aedes Aegypti Informs Arbovirus Vector Control', *Nature*, 563: 501–7.

McBride, C. S. et al. (2014) 'Evolution of Mosquito Preference for Humans Linked to an Odorant Receptor', *Nature*, 515: 222–7.

Mifsud, J. C. O. et al. (2022) 'Transcriptome Mining Extends the Host Range of the Flaviviridae to Non-bilaterians', *Virus Evolution*, 9: veac124.

Nayfach, S. et al. (2021) 'Metagenomic Compendium of 189,680 DNA Viruses from the Human Gut Microbiome', *Nature Microbiology*, 6: 960–70.

Nene, V. et al. (2007) 'Genome Sequence of Aedes Aegypti, A Major Arbovirus Vector', *Science (New York, N.Y.)*, 316: 1718–23.

Neri, U. et al. (2022) 'Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages', *Cell*, 185: 4023–4037.e18.

Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Palatini, U. et al. (2020) 'Improved Reference Genome of the Arboviral Vector Aedes Albopictus', *Genome Biology*, 21: 215.

Palatini U, Contreras C A, Gasmi L and Bonizzoni M 2022 Endogenous viral elements in mosquito genomes: current knowledge and outstanding questions Current Opinion in Insect Science 49 22–30

Peccoud, J. et al. (2018) 'A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated during Deep Sequencing Library Preparation', *G3: Genes Genomes Genetics*, 8: 1129–38.

Prachayangprecha, S. et al. (2014) 'Exploring the Potential of Next-Generation Sequencing in Detection of Respiratory Viruses', *Journal of Clinical Microbiology*, 52: 3722–30.

Presti, R. et al. (2009) 'Quaranfil, Johnston Atoll, and Lake Chad Viruses are Novel Members of the Family Orthomyxoviridae', *Journal of Virology*, 83: 11599–606.

Ramos-Nino, M. E. et al. (2020) 'Metagenomic Analysis of Aedes Aegypti and Culex Quinquefasciatus Mosquitoes from Grenada, West Indies', *PLoS One*, 15: e0231047.

Rozo-Lopez, P. et al. (2023) 'Untangling an Insect's Virome from Its Endogenous Viral Elements', *BMC Genomics*, 24: 636.

Russo, A. G. et al. (2019) 'Novel Insights into Endogenous RNA Viral Elements in Ixodes Scapularis and Other Arbovirus Vector Genomes', *Virus Evolution*, 5: vez010.

Shi, M. et al. (2016) 'Redefining the Invertebrate RNA Virosphere', *Nature*, 540: 539–43.

Shi, C. et al. (2019) 'Stable Distinct Core Eukaryotic Viromes in Different Mosquito Species from Guadeloupe, Using Single Mosquito Viral Metagenomics', *Microbiome*, 7: 121.

Simmonds, P. et al. (2017) 'Virus Taxonomy in the Age of Metagenomics', *Nature Reviews, Microbiology*, 15: 161–8.

Smits, S. L. et al. (2015) 'Recovering Full-length Viral Genomes from Metagenomes', *Frontiers in Microbiology*, 6: 1069.

Sofuku, K. et al. (2018) 'Influence of Endogenous Viral Sequences on Gene Expression', in *Gene Expression and Regulation in Mammalian Cells—Transcription from General Aspects*. IntechOpen.

Staginnus, C., and Richertpoggeler, K. R. (2006) 'Endogenous Pararetroviruses: Two-faced Travelers in the Plant Genome', *Trends in Plant Science*, 11: 485–91.

Sutton, T. D. S. et al. (2019) 'Choice of Assembly Software Has a Critical Impact on Virome Characterisation', *Microbiome*, 7: 12.

Tassetto, M. et al. (2019) 'Control of RNA Viruses in Mosquito Cells through the Acquisition of vDNA and Endogenous Viral Elements', *eLife*, 8: e41244.

Taylor, D. J., Leach, R. W., and Bruenn, J. (2010) 'Filoviruses are Ancient and Integrated into Mammalian Genomes', *BMC Evolutionary Biology*, 10: 193.

Ter Horst, A. M. et al. (2019) 'Endogenous Viral Elements are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs', *Journal of Virology*, 93: e02124–18.

Wallau, G. L. (2022) 'RNA Virus EVEs in Insect Genomes', *Current Opinion in Insect Science*, 49: 42–7.

Weiss, R. A. (2006) 'The Discovery of Endogenous Retroviruses', *Retrovirology*, 3: 67.

Whelan, S. P. J., Barr, J. N., and Wertz, G. W. (2004) 'Transcription and Replication of Nonsegmented Negative-Strand RNA Viruses', in Kawaoka, Y. (ed.) *Biology of Negative Strand RNA Viruses: The Power of Reverse Genetics*. pp. 61–119. Heidelberg, Germany: Springer.

Whickham, H. (2016) *Ggplot2: Elegant Graphics for Data Analysis*, 3rd edn. New York: Springer.

Whitfield, Z. J. et al. (2017) 'The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the Aedes Aegypti Genome', *Current Biology: CB*, 27: 3511–19.e7.

Wolf, Y. I. et al. (2020) 'Doubling of the Known Set of RNA Viruses by Metagenomic Analysis of an Aquatic Virome', *Nature Microbiology*, 5: 1262–70.

Yu, G. et al. (2018) 'Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using *Ggtree*', *Molecular Biology and Evolution*, 35: 3041–3.

Zhdanov, V. M. (1975) 'Integration of Viral Genomes', *Nature*, 256: 471–473.

Virus Evolution, 2024, **10(1)**, 1–14
DOI: https://doi.org/10.1093/ve/vead088
Advance Access Publication 28 December 2023

Research Article