Research Square

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

Genome-wide Characterization of Endogenous Viral Elements Reveals Extensive Viral Diversity in Culicinae Mosquitoes

Vikas Sharma

sharmavikas3529@gmail.com

ICMR-National Institute of Virology, Pune, India

Neha Verma

ICMR-National Institute of Virology, Pune, India

Ankush Sharma

ICMR-National Institute of Virology, Pune, India

Alfia Ashraf

ICMR-National Institute of Virology, Pune, India

Ranjana Raju

ICMR-National Institute of Virology, Pune, India https://orcid.org/0009-0001-0337-4177

Harsh Apurva

ICMR-National Institute of Virology, Pune, India

Kalichamy Alagarasu

ICMR-National Institute of Virology, Pune, India

Anisha Pulinchani

ICMR-National Institute of Virology, Pune, India

Surendra Kumar

ICMR-National Institute of Virology, Pune, India

Virendra Meena

ICMR-National Institute of Virology, Pune, India

Himanshu Kaushal

Indian Council of Medical Research-National Institute of Virology (ICMR-NIV), Pune, Maharashtra, India, Pin-411021 https://orcid.org/0000-0003-0935-9039

Naveen Kumar

National Institute of Virology https://orcid.org/0000-0003-3974-9409

Article

Keywords: Endogenous viral elements, Culicinae mosquitoes, Comparative genomics, Virus–host co-evolution, Retrovirus-like elements, Phylogenetic analysis

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10266412/v1

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

Additional Declarations: There is **NO** Competing Interest.

Genome-wide Characterization of Endogenous Viral Elements Reveals Extensive Viral Diversity in *Culicinae* Mosquitoes

Vikas Sharma[1,2]*, Neha Verma[1], Ankush Kumar Sharma[1], Alfia Fathima Ashraf[1], Ranjana Mariyam Raju[1], Harsh Apurva[1], Kalichamy Alagarasu[1,2], Anisha Pulinchani[1,2], Surendra Kumar[1], Virendra Kumar Meena[1,2], Himanshu Kaushal[1,2], and Naveen Kumar[1,2]

[1]ICMR-National Institute of Virology, Pune, India

[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

*First Author and Corresponding author

Address for correspondence
Dr. Vikas Sharma,
Scientist-C,
ICMR-National Institute of Virology,
20-A, Dr. Ambedkar Road, Pune,
Email: sharmavikas3529@gmail.com; sharmavikas.niv@icmr.gov.in

Abstract

Endogenous viral elements (EVEs) are viral sequences integrated into host genomes that shape genome evolution, antiviral immunity, and gene regulation; however, their diversity and evolutionary significance in mosquitoes remain insufficiently characterized. A genome-wide analysis of EVEs was conducted across 19 RefSeq Culicinae mosquito genomes, identifying 77,141 putative EVEs representing 1,538 viral species, although these elements occupy less than 1% of mosquito genomes. EVE abundance varied significantly among species, with *Aedes koreicus* exhibiting the highest EVE burden. Approximately 62% of EVEs were homologous to Trichoplusia ni TED-like and *Cladosporium fulvum* T1-like retroviral elements. EVEs demonstrated non-random chromosomal distributions, integration hotspots, and frequent associations with transposable element-related domains. Phylogenetic analyses revealed multiple previously unrecognized EVE clades, while transcriptomic and PCR analyses confirmed the expression and conservation of dominant EVEs in Indian *Aedes aegypti*. This study provides the most comprehensive catalog of mosquito EVEs to date and establishes a foundation for understanding their roles in mosquito evolution, antiviral immunity, and vector biology.

Keywords: Endogenous viral elements; *Culicinae* mosquitoes; Comparative genomics; Virus–host co-evolution; Retrovirus-like elements; Phylogenetic analysis

Introduction

The family *Culicidae* represents a diverse and evolutionarily ancient group of mosquitoes, encompassing roughly 4,500 species split into the subfamilies *Anophelinae* and *Culicinae*. These two subfamilies diverged approximately 180–250 million years ago, resulting in notable differences in genome architecture. Notably, *Culicinae* mosquitoes possess much larger genomes, primarily due to an expansion of transposable elements (TEs) [1–4]. As major arthropod vectors, *Culicinae* mosquitoes transmit a broad range of human and animal pathogens including viruses, protozoa, and nematodes that cause significant diseases such as yellow fever, chikungunya, dengue, Zika, and malaria [5,6].

Despite frequent exposure to arboviruses, mosquitoes seldom display overt disease symptoms, highlighting their remarkable tolerance to persistent viral infections [7,8]. This tolerance is largely attributed to small RNA-based antiviral pathways, particularly RNA interference (RNAi) and PIWI-interacting RNA (piRNA) systems, which regulate viral replication and influence long-term virus–host co-evolution [9,10]. Previously published studies indicate that endogenous viral elements (EVEs), which are DNA- or RNA-derived viral sequences integrated into mosquito host genomes, may play an important role in these antiviral mechanisms [11–14].

EVEs are pervasive components of eukaryotic genomes, including those of humans, insects, and plants, where they contribute to genome evolution, antiviral immunity, gene regulation, and the emergence of evolutionary innovations [15–20]. The study of these ancient viral remnants and their evolutionary impact is known as paleovirology [21,22]. EVEs originate through two principal mechanisms of genomic integration. Retroviruses integrate into germline DNA through a virus-encoded integrase, whereas non-retroviral viruses are thought to become endogenized through host-mediated mechanisms, including retrotransposon-driven reverse transcription and non-homologous recombination during DNA double-strand break repair [17,23,24]. These endogenous remnants of both retroviral and non-retroviral origins provide valuable insights into long-term host–virus co-evolution, adaptation, and genome evolution [16,21,25].

In mosquitoes, evidence suggests that retrotransposon-mediated integration is the primary mechanism underlying EVE formation [11,26]. The abundance of EVEs correlates with the diversity and activity of transposable elements (TEs), and EVEs are frequently found in TE-rich genomic regions and piRNA clusters [10,27]. Notably, EVE integrations are not randomly distributed within mosquito genomes but are enriched in piRNA clusters, where they generate antisense piRNAs that target cognate viral RNAs [11,27]. Many EVEs are also located adjacent to long terminal repeat (LTR) retrotransposons, supporting a role for transposable elements in the integration, maintenance, and diversification of viral sequences [10,11]. However, EVEs outside piRNA clusters may retain intact viral open reading frames and remain transcriptionally active, indicating functional diversity. Although most EVEs are highly fragmented, they can still generate virus-derived small RNAs that contribute to sequence-specific antiviral immunity [28,29].

Despite these advances, our understanding of mosquito EVEs remains fragmentary. Previous studies have focused primarily on EVEs derived from non-retroviral RNA viruses in a limited number of mosquito species, revealing substantial variation in EVE abundance, genomic distribution, and transcriptional activity [11,30–32]. Most characterized mosquito EVEs are partial or fragmented viral sequences, consistent with their ancient origin and progressive sequence degeneration following genomic integration [11,33]. However, the overall viral diversity of EVEs across *Culicinae* mosquitoes, particularly retrovirus-like elements, their genome-wide organization, evolutionary origins, and associations with transposable elements and antiviral pathways remain poorly understood.

In this study, we conducted a genome-wide survey of 19 RefSeq *Culicinae* mosquito genomes to characterize the diversity, abundance, and genomic organization of EVEs. Through comparative genomics, protein domain annotation, synteny analysis, phylogenetic reconstruction, transcriptomic evidence, and experimental validation, we present a comprehensive overview of the EVE landscape in *Culicinae* mosquitoes. Our results reveal extensive EVE diversity, lineage-specific genomic patterns, strong associations with TEs, and multiple novel viral lineages. We further confirm the genomic integration of selected *Trichoplusia ni TED-like* (*TED*) and *Cladosporium fulvum T1-like* (*Clado*) EVEs in Indian Aedes aegypti using PCR and Sanger sequencing. Collectively, these findings establish the first comprehensive comparative

framework for understanding the origins, evolution, and genomic landscape of EVEs in Culicinae mosquitoes, providing a valuable resource for future research into their roles in antiviral immunity, vector biology, and host–virus co-evolution.

Methods

Data Collection and EVE Detection Pipeline

Nineteen complete RefSeq genomes and twenty-eight assembled transcriptomes from *Culicinae* mosquitoes were downloaded from the NCBI GenBank FTP repository on March 6, 2025 (Supplementary Table 1: Sheet 1, Supplementary Table 2: Sheet 1). These datasets were screened to identify EVEs or viral homologs present within mosquito genomes and transcriptomes using *EEFinder (version 1.0)*, an automated pipeline for EVE detection and classification [34]. The pipeline employed BLASTX searches[35] against the NCBI Viral RefSeq Protein Database (default parameters) to detect viral homologs embedded within host sequences. To minimize false positives, experimentally or functionally characterized host protein sequences from multiple mosquito species were used as bait sequences in the *EEFinder* workflow [34]. Output tables were merged to remove overlapping regions, and only candidate proteins showing at least 30% sequence similarity to viral proteins were retained for downstream analyses. Viral-like homologs identified in this manner were further processed and visualized using custom R scripts (R v 4.6.0).

To assess the prevalence of these EVEs in Indian mosquito populations, two *Aedes aegypti* genomes from India were obtained from NCBI GenBank and examined for their presence using the *EEFinder* pipeline (Supplementary Table 1: Sheet 1). The resulting data were analyzed using custom R scripts.

Phylogenetic and Clustering-Based Analysis

The two most abundant viral groups identified, *Trichoplusia ni TED virus* (*TED*-like) and *Cladosporium fulvum T-1 virus* (*Clado*-like) viruses, were selected for phylogenetic and sequence clustering analyses. Genomic loci corresponding to these viruses were translated into protein sequences using Prodigal v2.6.3 [36]. Sequences shorter than 90 amino acids were excluded to minimize the inclusion of incomplete or low-information sequences using SeqKit v2.13.0 [37]. Protein sequences from each mosquito genome were clustered using CD-HIT v 4.8.1 [38] at 60% sequence identity to remove redundancy and obtain representative sequences. Representative sequences from all mosquito genomes were combined for further analysis. The longest EVE representative was used to identify a reference dataset via BLASTP v2.17.0+ [39]. This reference set was clustered with CD-HIT v4.8.1 [38] at 60% identity and used for phylogenetic analysis. This combined dataset of representative EVE and viral reference proteins was used for both phylogenetic reconstruction and t-distributed stochastic neighbor embedding (t-SNE) analyses. For phylogenetic analysis, sequence length statistics were calculated, and sequences below the

average length were excluded with SeqKit v2.13.0 [37]. Multiple sequence alignments were created with MAFFT (v7.490)[40]. Sequence alignment was automatically curated using MaxAlign[41] and trimAl v1.4.rev15 [42]. Maximum-likelihood phylogenetic trees with optimal amino acid substitution models were generated using IQ-TREE 2 v1.6.12[43]. Trees were visualized and annotated using the *ggtree* (v4.2.0) package in R[44].

Therefore, t-SNE [45,46] was performed as a complementary approach to investigate sequence-level relationships among representative EVEs and their viral homologs. Separate analyses were conducted for the *TED*-like ORF-B and *Clado*-like reverse transcriptase (RT) protein datasets. High-dimensional sequence features were projected into a two-dimensional space using the standard t-SNE algorithm for visualization.

Density Distribution of EVE Copy Numbers across Mosquito Genomes

To analyze the genomic distribution of EVEs, only mosquito genomes with chromosome-level assemblies were used. Chromosomes were partitioned into non-overlapping 10-kb bins, and the number of EVE sequences (ORF-B for *TED* and RT for *Clado*-like virus) in each bin was calculated. To account for differences in chromosome length, EVE counts were normalized by the total chromosome length. Normalized density values were used to assess distribution patterns, and the results were visualized using custom R scripts.

EVE Integration Site Functional Annotation and synteny analysis

For detailed analysis, the two most abundant viral species were selected. For each putative EVE locus, 5 kb of upstream and downstream genomic regions were extracted with custom Python scripts (v 3.13.5). Functional annotation of these flanking regions was performed using RPS-BLAST [39] against the NCBI Conserved Domain Database (CDD)[47]. Detected protein domains were summarized to identify patterns of functional enrichment near EVE sites. The most enriched domains underwent Gene Ontology (GO) annotation analysis using PFamtoGO [48], and results were visualized to provide an overview of domain enrichment. Protein domains identified within the ±5 kb regions surrounding EVE loci were compiled and analyzed using R (v4.6.0). Syntenic organization display analyses, conserved Pfam domains were grouped according to their corresponding viral ORFs: ORF-A = Pfam03732; ORF-B (Pol) = reverse transcriptase (Pfam00078), retroviral aspartyl protease (Pfam00665), RNase H (Pfam17917), and integrase zinc-binding (Pfam17921); and ORF-C (Env-like) = Baculo_F (Pfam12259).

Analysis of PIWI and defense system Proteins in the 5kb flanking region

PIWI-like proteins and piRNAs play a key role in mosquito antiviral defense by suppressing viral expression [49]. For this analysis, five Pfam protein domains: piwi (Pfam 02171), PAZ (Pfam 02170), AgroN (Pfam 16486), AgroL1 (Pfam 08699), and ArgoL2 (Pfam 16488)—were selected based on structural completeness and previous literature[50] . The presence of these domains was screened within the 5 kb flanking regions of *TED*-like and *Clado*-like EVEs across all 19 mosquito genomes. To further assess the genomic context of EVE integration sites, protein

domains associated with known defense systems were identified by comparing the detected Pfam annotations with the curated Pfam reference list provided by the DefenseFinder database (https://defensefinder.mdmlab.fr/wiki/defense-systems)[51]. Domains matching this curated reference were classified as putative defense-related proteins. The distribution and abundance of PIWI-associated and defense-related domains surrounding EVE loci were subsequently evaluated to determine whether these genomic regions exhibit enrichment for host defense functions. Domain abundance matrices were generated in R (v4.6.0) from the occurrence of Pfam domains across mosquito species. The 20 most abundant domains were visualized using bubble plots generated with ggplot2 (v4.0.3), in which bubble size and color intensity represented log-transformed domain counts.

Experimental Validation of EVEs Presence in *Aedes aegypti*

Experimental validation was performed using Indian *Aedes aegypt*i genomic DNA, as this species was available for in-house molecular validation. The described experimental workflow can be readily applied to other mosquito species.

Primer Design

Primers were designed for the predicted EVEs detected in the Indian Aedes aegypti genomes. From these genomes, EVEs were extracted using a custom Python script, and the retrieved sequences were clustered using CD-HIT v4.8.1[38] at 99% identity, with a minimum 90% overlap for shorter sequences to reduce redundancy. Representative clusters were selected based on larger cluster size or longer EVE lengths. Cluster information was extracted and sorted using AWK scripts. Multi-FASTA clusters were submitted to the NCBI primer design tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/), with the "Primer common for group of sequences" option. Primer specificity checks excluded Homo sapiens, Aedes aegypti, and Aedes albopictus genomes, though some overlap with Aedes aegypti was unavoidable due to EVE fragmentation in the genome.

PCR Amplification and Product Visualization

PCR reactions (20 µL) contained 10 µL EmeraldAmp® GT PCR Master Mix (2x; Takara, Cat No.: RR310), 1 µL of each primer (0.5 µM), 4.3 µL mosquito genomic DNA (~50 ng), and 3.7 µL nuclease-free water. A Touchdown PCR was performed in a 20 µL reaction volume. The cycling protocol consisted of an initial denaturation at 96 °C for 3 min, followed by 10 touchdown cycles of 94 °C for 30 s, annealing from 60 °C to 55 °C with a decrement of 0.5 °C per cycle for 30 s, and extension at 72 °C for 30 s. This was followed by 30 additional cycles of 94 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 30 s. A final extension was carried out at 72 °C for 5 min before holding at 4 °C.

PCR products were visualized by 2% agarose gel electrophoresis in 1X TAE buffer, stained with Nancy-520 (SKU: 01494-500UL), with 100 bp DNA marker (BR Biochem Cat No.: BM001R500) and run at 90 V. Amplicons of expected size were purified directly from PCR

product when single band was observed; In case multiple bands appeared then amplicon corresponding to the expected size were excised and purified using the QIAquick Gel and PCR Extraction Kit (QIAGEN, Cat No.: 28506). Purified DNA was quantified and, if necessary, re-amplified for Sanger sequencing.

Sanger sequencing and sequence validation

Sequencing reactions (10µL) included ~60ng purified product, 0.5µM primer, was carried out using Applied Biosystems sequencing PCR kit, 1µL BigDye Terminator v3.0 REF 4336909), 1.5µL sequencing buffer, and water. Reactions underwent denaturation at 94°C, followed by 35 cycles of 96°C, 50°C, and 60°C steps. Sequencing was performed on in-house Applied Biosystems 3100 Avant Genetic Analyzer.

Forward and reverse chromatograms were manually inspected in UGENE v51.0 [52]. Low-quality terminal regions were trimmed, and high-quality sequences were exported as FASTA files. Retained reverse reads were converted to reverse complements using a custom command-line script. Forward and reverse reads were independently queried against the NCBI viral nr protein database (downloaded 16 May 2025) using BLASTX[35] (E-value = 1×10^{-3}, BLOSUM45, SEG disabled, max_target_seqs = 10, max_hsps = 1) to assess recovery of the expected viral lineage. The highest-quality forward and reverse-complement reads were then pairwise aligned using MUSCLE v5.3 (command-line implementation)[53]. The resulting alignments were manually inspected in UGENE v51.0[52], and the concordant overlapping region was manually extracted. Alignment gaps were subsequently removed using a custom command-line script. Concordant sequences were then queried against the same database using a more permissive BLASTX[35] search (E-value = 10, SEG disabled, max_target_seqs = 20, max_hsps = 1) to confirm viral homology and assess retention of the expected viral lineage.

Results

Distribution and Taxonomic Diversity of Identified EVEs across 19 Mosquito Genomes

To explore the prevalence, diversity, and taxonomic composition of EVEs in mosquito genomes, we analyzed 19 complete RefSeq genomes from the Culicinae subfamily using the computational pipeline EEfinder. This approach, which incorporates BLASTX similarity searches, identified 77,141 potential EVEs that exhibit at least 30% amino acid identity to known viruses. These EVEs correspond to 1,538 distinct viral species, 268 genera, and 67 viral families that have become integrated into the mosquito genomes we studied (Supplementary Table 1, Sheets 1 and 2). Quantification of the genomic composition revealed that single-stranded RNA viruses with reverse transcriptase (ssRNA–RT) were the most prevalent, accounting for 68% of all detected EVEs (Fig. 1a). This finding highlights retrovirus-like elements as the major source of viral integrations in

mosquito genomes. In comparison, double-stranded DNA (dsDNA) viruses comprised 25.6%, double-stranded DNA viruses with reverse transcriptase (dsDNA–RT) 4%, and negative-sense single-stranded RNA viruses (ssRNA−) 1.7%. Minor components included positive-sense single-stranded RNA viruses (ssRNA+) (0.1%), generic RNA viruses (0.2%), double-stranded RNA viruses (dsRNA) (0.2%), and unclassified viral types (0.2%) (Fig. 1a). These results indicate that most EVEs in mosquito genomes originate from retrovirus-like elements, underscoring their long-term integration and persistence within host genomes.

Taxonomic profiling of the 77,141 putative EVEs demonstrated a broad distribution across viral families. *Metaviridae* was the most abundant family, representing 62% of all identified elements, followed by *Polydnaviriformidae* (6.1%), *Retroviridae* (5.6%), *Poxviridae* (5.3%), and *Iridoviridae* (4.2%). The remaining 0.5% were distributed among several minor viral families (Fig. 1b). At the genus level, *Errantivirus* accounted for 52.1% of all EVEs, with *Metavirus* (10.3%), *Bracoviriform* (5.9%), and *Deltaentomopoxvirus* (4.1%) also prominent. The remaining 9.6% were attributed to other genera (Fig. 1c). At the species level, *TED-like virus* was the most frequently detected (52.1% of total EVEs), followed by *Clado-like virus* (10.3%) and *Melanoplus sanguinipes entomopoxvirus* (4.1%).The remaining 19.4% corresponded to additional viral species (Fig. 1d). Analysis of host associations of EVEs revealed a strong association with insect-associated viruses, which accounted for 53.2% of all EVEs, reflecting the insect-specific evolutionary ecology of mosquito genomes. A considerable proportion (29.2%) could not be assigned to a host due to incomplete metadata in reference databases. Of the remaining annotated EVEs, smaller proportions were linked to protists (3.9%), fish (2.9%), mammals (2.4%), and plants (1.5%), with an additional 6% classified as other hosts. Only minor fractions were associated with humans (0.4%), bacteria (0.3%), arthropods (0.1%), and fungi (0.006%). Altogether, these findings suggest that mosquito EVEs predominantly originate from insect-associated and retrovirus-like viruses, while also capturing evidence of cross-kingdom viral diversity. This pattern highlights the broad ecological exposure of mosquitoes to diverse viral communities and the long-term accumulation of viral sequences within their genomes which may influence mosquito biology and vector competence.

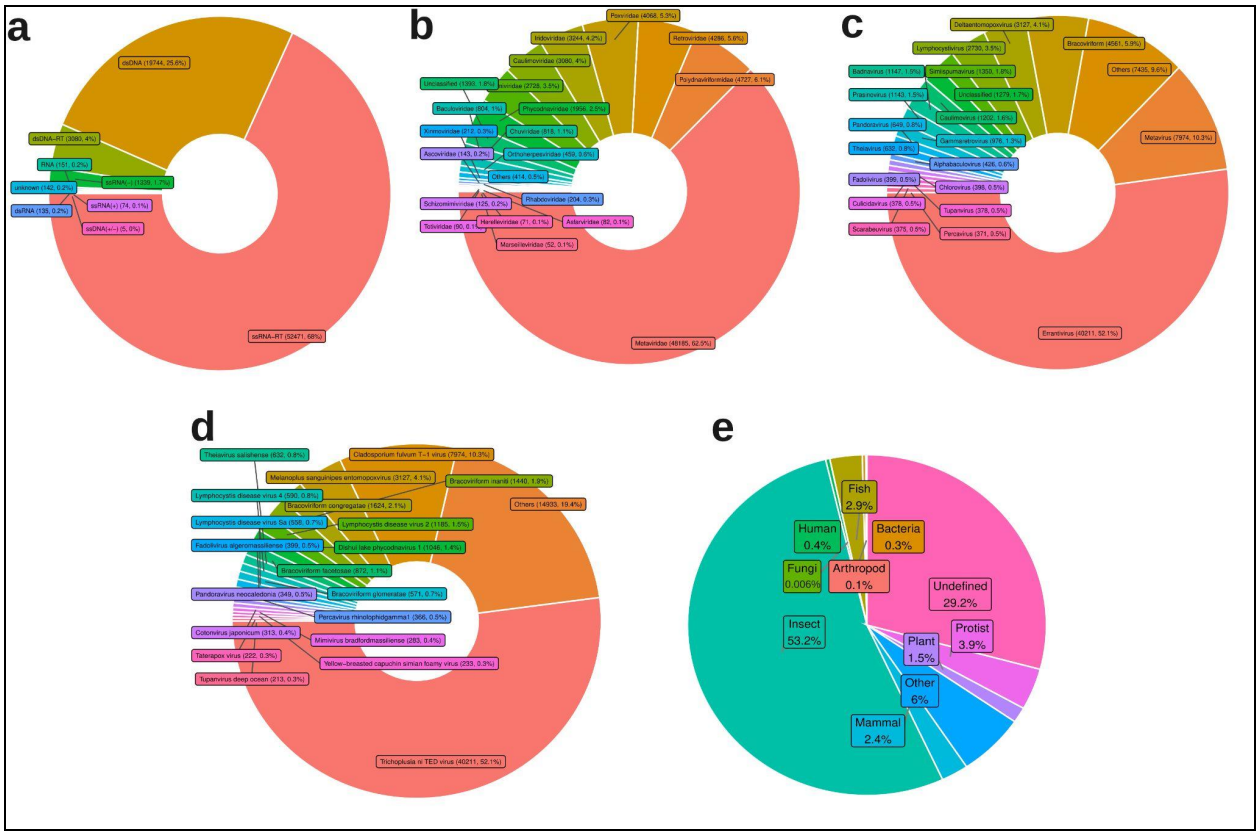


Fig. 1: Diversity of putative EVEs identified in 19 mosquito genomes. (a) Molecular type of putative EVEs **(b),** Diversity at family level **(c),** Diversity at Genus level **(d),** Diversity at Species level **(e),** Host diversity.

Distribution Patterns of EVEs across Mosquito Species

Having established the diversity and taxonomic breadth of EVEs in mosquito genomes, we next examined their distribution across different mosquito species. The abundance of identified EVEs varied substantially among mosquito species, ranging from 676 elements in *Culex pipiens pallens* to 10,446 in *Aedes japonicus*. EVE abundance was normalized by dividing the number of viral homologs by the genome size (in megabases, Mb). This analysis revealed that *Aedes koreicus* exhibited the highest EVE density (9.4 per Mb), followed by *Ochlerotatus camptorhynchus* (7.7), *Aedes japonicus* (7.5), and *Armigeres subalbatus* (5.4). In contrast, the lowest density was observed in *Culex pipiens pallens* (1.1 per Mb) (Fig. 2a).

Notably, while *Aedes japonicus* contained the highest absolute number of EVEs, *Aedes koreicus* showed the highest relative density, highlighting the distinction between total abundance and genome-normalized enrichment. Consistent with this trend, EVEs occupied less than 1% of the host genome in all species.The highest genomic proportion was observed in Aedes koreicus

(0.53%), followed by Ochlerotatus camptorhynchus (0.49%) and Aedes aegypti (0.43%), whereas Culex quinquefasciatus exhibited the lowest proportion (0.12%) (Fig. 2b). Together, these results showed lineage-specific diversification, accumulation, and retention of EVEs in Culicinae mosquitoes.

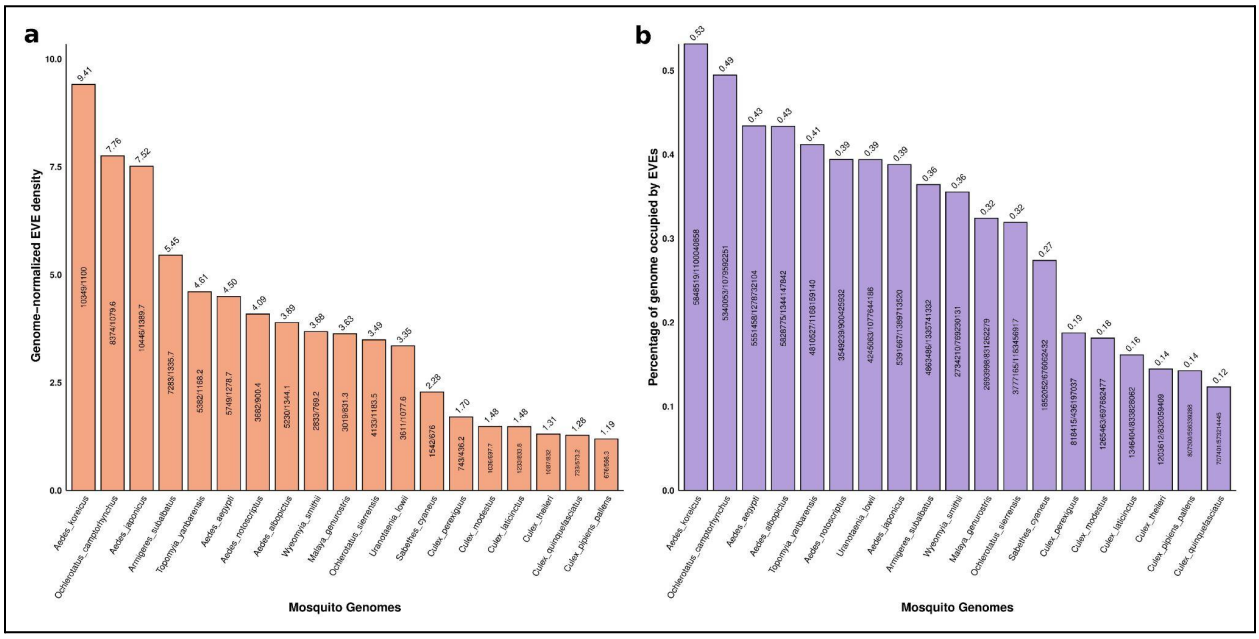


Fig. 2: **Distribution of all identified EVEs in the Mosquito genome.** (a) Representing the overall normalized EVEs density observed per MB genome. (b) Percentage of the genome occupied by EVEs for each species.

Validation of EVE Profiles in Indian Mosquito Genomes

To validate the broader applicability and stability of EVE profiles, we extended our analysis to recently sequenced Indian mosquito genomes. Since, the initial screening was performed on 19 high-quality RefSeq mosquito genomes that did not include Indian isolates. To address EVEs diversity and evolution in the context of Indian mosquito species we screened two recently sequenced Indian *Aedes aegypti* mosquito genomes (GCA_009613055.1_IBAB_Aaeg_KPA_1.0, GCA_009613065.1_IBAB_Aaeg_KOR_1.0_genomic) to validate the presence and consistency of EVE signatures. In the Indian *Aedes aegypti* genome, 11,374 EVEs were identified (Supplementary Table 1, Sheet 3). Analysis of the top 20 viral species revealed that the most abundant elements were *TED-like virus (6,995 copies), Clado-like Virus (1,998), Melanoplus sanguinipes entomopoxvirus (380), Dishui Lake phycodnavirus 1 (297),* and *Bracoviriform inaniti (158)* (Supplementary Fig. 1). The abundance patterns of the four most dominant viral species closely matched those observed in the 19 reference mosquito genomes, indicating a similar EVE composition across geographically distinct mosquito populations. Several additional EVEs were uniquely detected in the Indian dataset, including *Gylbovirus aagae* (104),

Caulimovirus venafragariae (30), *Ohlsrhavirus ohlsdorf* (22), *Chrysochromulina ericina virus* (21), *Gambievirus senegalense* (21), *Menghai flavivirus* (20), and *Paramecium bursaria Chlorella virus NY2A* (20). These region-specific elements likely reflect differences in ecological exposure and local viral diversity. Collectively, these findings indicate that although geographic variation influences the presence of certain EVEs, the overall EVE landscape remains similar across mosquito populations. The widespread presence of these viral elements across all analyzed genomes suggests that common integration events likely occurred in the ancestral lineage of mosquitoes, resulting in ancient, evolutionarily stable viral integrations.

Landscape of EVE Diversity across Studied Mosquito Genomes

Building on the validation results, we next assessed the diversity of EVEs across the studied mosquito genomes in greater detail. EVE diversity varied markedly among analyzed mosquito species (Supplementary Fig. 2). The number of viral families represented by EVEs ranged from 9 to 47, with genus-level diversity from 22 to 172 and species-level diversity from 38 to 772. *Aedes koreicus* exhibited the highest family-level diversity (47 viral families), while *Aedes japonicus* displayed the greatest genus and species-level diversity (172 genera, 772 species). By contrast, *Culex quinquefasciatus* showed the lowest diversity across all taxonomic ranks (9 families, 24 genera, 38 species). When species were arranged by taxonomic relationships, lineage-associated patterns emerged. Members of the *Aedes/Ochlerotatus* lineage generally exhibited greater EVE diversity than most *Culex* species, particularly at genus and species levels. Notably, *Aedes japonicus, Aedes koreicus*, and *Ochlerotatus camptorhynchus* harbored especially diverse viral repertoires, whereas most *Culex* species displayed lower richness. Substantial variation among closely related taxa indicates that EVE diversity is not strictly determined by taxonomy.

Comparison with genome assembly size revealed no consistent correlation: although some species with large genomes (e.g., *Aedes japonicus, Aedes koreicus*) exhibited high EVE diversity, others (e.g., *Topomyia yanbarensis*, 1,168 Mb) did not. Conversely, species with smaller genomes (e.g., *Armigeres subalbatus, Aedes aegypti*) sometimes harbored comparatively diverse repertoires. Thus, genome size is insufficient to explain variation in EVE richness. Collectively, these results reveal highly heterogeneous EVE distributions across *Culicinae* genomes, concentrated within particular lineages and independent of genome size, suggesting lineage-specific evolutionary processes may shape EVE diversity.

Comparative Distribution of *Trichoplusia ni TED* -like and *Cladosporium fulvum* T-1–like EVEs in Mosquito Genomes

Given the observed dominance of retrovirus-like EVEs, we directed our subsequent analyses toward the two most abundant virus-derived EVE groups identified across mosquito genomes. Among all detected EVEs, single-stranded RNA viruses with reverse transcriptase (ssRNA–RT) were the most prevalent, followed by double-stranded DNA (dsDNA), dsDNA–RT, and other viral categories. ssRNA–RT viruses were consistently identified in all mosquito genomes,

indicating their widespread distribution. To further characterize this dominant class, we focused in detail on the two most abundant retrovirus-like EVEs: those homologous to *TED*-like EVEs (40,211 copies; ≥30% similarity to YP_009507248.1) and *Clado*-Like EVEs (7,973 copies; ≥30% similarity to YP_009666308.1) (Supplementary Table. 2).

For *TED*-like EVEs, a total of 40,211 copies were identified across ORF A and ORF B regions, with the vast majority (40,056) corresponding to ORF B. Accordingly, all subsequent analyses were specifically focused on the ORF B protein of *TED*-like EVEs. Among mosquito genomes, *Armigeres subalbatus* had the highest abundance (4,352 copies) and the highest normalized density (3.26 per Mb), while Culex pipiens pallens had the lowest representation, with 444 copies and a density of 0.78 per Mb (Supplementary Table. 1, Sheet2; Fig. 3a).

Similarly, 7,973 copies of *Clado*–like EVEs were identified across all genomes. For downstream analysis, we specifically focused on the reverse transcriptase (RT) domain of *Clado*–like EVEs. The highest abundance was found in *Uranotaenia lowii* (840 copies; density = 0.78 per Mb), and the lowest in *Culex pipiens pallens* (90 copies; density = 0.16 per Mb) (Supplementary Table 1, Sheet 2; Fig. 3b).

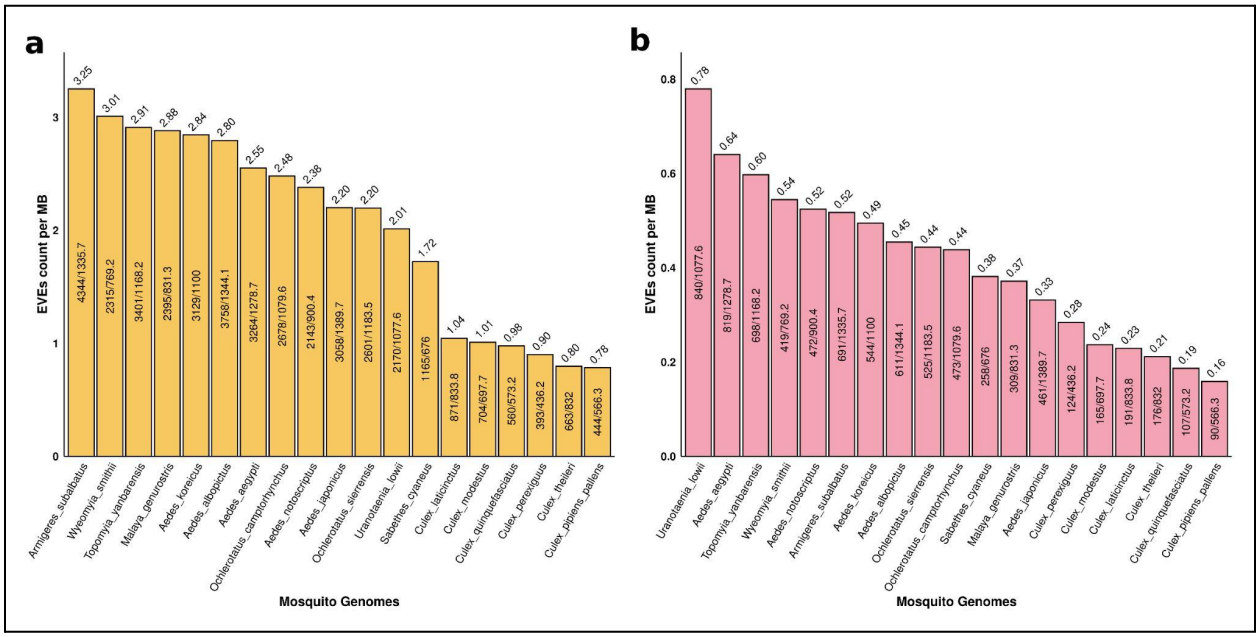


Fig. 3: **EVEs distribution bar plots based on normalized values of the top two Viral Species. (A)** EVEs density per MB genome of *TED-like virus*. **(B)** EVEs density per MB genome of *Clado-like* virus.

Comparative Evolutionary Analysis of *TED-like* and *Clado*-like EVEs

To investigate the evolutionary dynamics of the most abundant EVEs, we analyzed both sequence-length variation and phylogenetic relationships among *TED*-like and *Clado*-like viral elements across mosquito genomes. A total of 40,211 nucleotide sequences were translated into 84,513 amino acid sequences and compared to reference viral ORFs. Most identified EVEs were substantially shorter than their corresponding full-length viral proteins, indicating they are predominantly fragmented. For *TED*-like elements, the reference ORF B (1,084 amino acids) was much longer than the majority of detected EVEs, which were mostly below the mean length of 274 amino acids (Fig. 4a). These results suggest *TED*-like EVEs are likely degraded remnants of ancient viral integrations rather than intact, functional elements.

A similar pattern was observed for *Clado*–like elements: 7,973 EVEs were translated into 8,855 amino acid sequences, and the reference reverse transcriptase ORF (1,045 amino acids) was longer than most corresponding EVEs, the majority of which fell below the mean length of 415 amino acids (Fig. 4b). Compared with TED-like elements, EVEs tended to be longer, suggesting fragmentation of EVEs is a common evolutionary process in the mosquito genome; however, extent of evolutionary decay and retention dynamics between the EVEs lineage may differ.

Clustering and Phylogenetic Classification of *TED* and *Clado*-like Viral Homologs

To further investigate the evolutionary origins and relationships of the most abundant EVEs identified in mosquito genomes, we performed clustering and phylogenetic analyses of *TED*-like and *Clado*-like viral homologs. For the TED-like dataset, 40,211 nucleotide sequences showing >30% similarity to the ORF-B protein of *Trichoplusia ni* TED virus (YP_009507247.1) were translated into 47,213 protein sequences. After removing sequences shorter than 90 amino acids, 37,251 proteins were retained for downstream analyses. To reduce redundancy, sequences were clustered at 60% identity, yielding 3,521 representative EVE proteins. These were analyzed alongside 27 reference viral proteins, yielding a dataset of 3,548 sequences. Following sequence preprocessing and quality filtering, 3,540 sequences remained for clustering analysis.

DBSCAN clustering of the t-SNE projection partitioned the dataset into 53 clusters (Supplementary Table 3, Sheet 1). Most clusters consisted exclusively of EVE sequences, highlighting the extensive diversity and abundance of TED-like EVEs (Supplementary Fig. 3a). Only four clusters (7.5%; 4/53) contained both EVEs and reference viral proteins, comprising 1,111 of the 3,521 representative EVEs (31.6%) and 24 of the 27 reference proteins. These results indicate that a subset of *TED*-like EVEs retains detectable sequence-feature similarity to known viral proteins, whereas the majority occupy distinct regions of sequence space and exhibit substantial divergence from currently characterized viral lineages.

To assess whether these clustering patterns reflected deeper evolutionary relationships, representative *TED*-like EVEs were subjected to phylogenetic analysis. To minimize the effects of sequence fragmentation, only sequences equal to or longer than the average EVE length (~304 amino acids) were retained. Following alignment with MAFFT and curation using MaxAlign, 882 high-quality EVE sequences were analyzed together with 19 reference retroviral proteins (Fig. 4c). The resulting maximum-likelihood phylogeny revealed extensive previously

uncharacterized diversity. Only 34 EVEs (3.8%) were positioned within clades containing reference viral sequences, whereas the overwhelming majority formed distinct, well-supported clades separated from known viral lineages by long branches. These findings suggest substantial evolutionary divergence following long-term persistence in host genomes and indicate that most *TED*-like EVEs are ancient endogenous viral lineages that are only distantly related to currently recognized retroviral groups.

For the *Clado*-like dataset, 7,973 EVE loci were translated into 8,854 protein sequences, of which 7,370 sequences longer than 90 amino acids were retained. Clustering at 60% sequence identity reduced the dataset to 1,237 representative proteins, which were analyzed together with 44 reference viral proteins, yielding 1,281 input sequences (Supplementary Table 3, Sheet 2). After sequence preprocessing and quality filtering, 1,280 sequences were retained for clustering analysis. DBSCAN clustering of the t-SNE projection classified these sequences into 40 groups, comprising 39 multi-sequence clusters (clusters 0–38) and one singleton outlier group (cluster -1) (Supplementary Table 3, Sheet 2; Supplementary Fig. 3b). Five clusters (12.8%; 5/39) contained both EVEs and reference viral proteins, including 693 of the 1,237 representative EVEs (56.0%) and 37 reference proteins. The remaining clusters were composed predominantly of EVEs, further demonstrating extensive sequence diversity and divergence from currently characterized viral lineages.

For phylogenetic reconstruction, only sequences equal to or longer than the average EVE length (~334 amino acids) were retained to minimize the influence of fragmention. After MAFFT alignment and MaxAlign curation, 284 high-quality EVE sequences were analyzed alongside 23 reference retroviral proteins, resulting in a final dataset of 307 sequences. Consistent with the clustering analysis, the maximum-likelihood phylogeny showed that most *Clado*-like EVEs formed distinct, well-supported clades separate from reference viral lineages, with no EVE sequences directly nested within known viral groups (Fig. 4d). Nevertheless, the overall phylogenetic placement of several EVE clades remained consistent with relationships to members of the *Metaviridae* and *Errantivirus* lineages, supporting their viral ancestry. Collectively, these results indicate that both *TED*-like and *Clado*-like EVEs comprise highly diverse assemblages of endogenous viral sequences that have undergone extensive divergence following integration into mosquito genomes, while retaining detectable evolutionary links to recognized retroviral lineages.

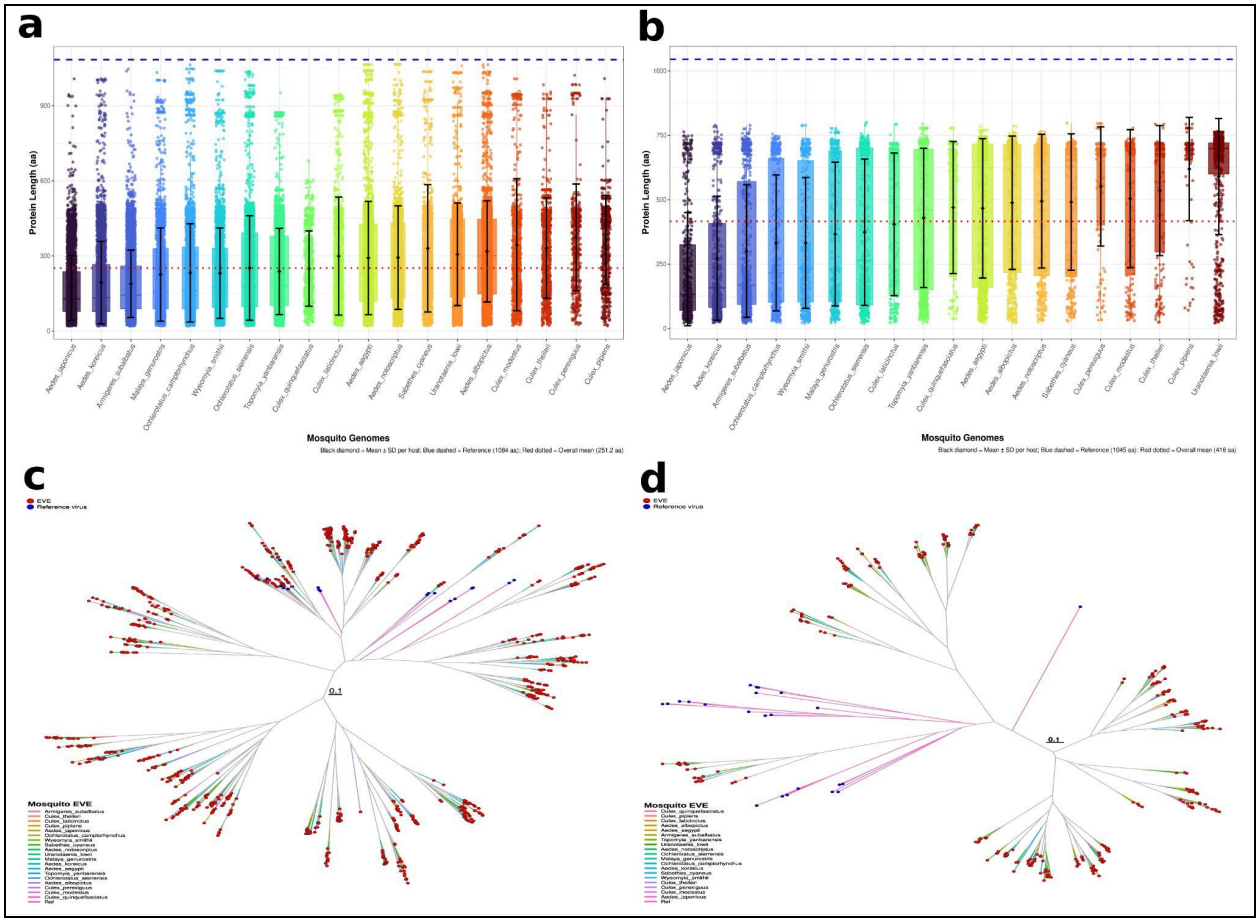


Fig 4: Sequence Length Variation and Phylogenetic Relationships of *TED*-like and *Clado*-like EVEs in Mosquito Genomes. (A) Sequence length distributions of *TED*-like virus ORF-B and identified EVEs across mosquito genomes. Boxplots show the EVE sequence length distributions for each mosquito species. The blue dashed line marks the reference viral ORF-B sequence (3,255 bp; YP_009507248.1), and the red dotted line shows the overall mean EVE sequence length (944 bp). (B) Sequence length distributions of *Clado*-like virus EVEs across mosquito genomes, with the blue dashed line indicating the reference viral sequence (3,138 bp; YP_009666308.1) and the red dotted line the overall mean EVE sequence length (1,465 bp). (C) Unrooted maximum-likelihood tree for *TED*-like viral proteins based on sequences ≥304 amino acids. (D) Unrooted maximum-likelihood tree for *Clado*-like viral proteins based on sequences ≥334 amino acids. For both trees, branch colors denote mosquito host lineage, red circles indicate EVEs from mosquito genomes, and blue circles indicate reference viral sequences. Scale bars represent amino acid substitutions per site; bootstrap support values are provided in Supplementary Figures S1 & S2.

Non-Random Chromosomal Distribution of EVEs (*ORF-B/RT homologs*) in Mosquito Genomes

The genomic distribution of ORF-B and RT homologs was analyzed across 16 mosquito species with chromosome-level assemblies. Chromosomes were segmented into 10 kb bins, and EVE counts were normalized by bin and chromosome length to enable intra- and interspecies

comparisons. To facilitate visualization and comparison across samples, the RT percentages were normalized using min–max scaling to a range of 1–100.

Both ORF-B (*TED*-like) and RT (*Clado*-like) homologs exhibited non-random distributions, characterized by pronounced hotspot formation across the chromosomal landscape, with enrichment often observed in specific chromosomal regions (Fig. 5a & 5b). This pattern was broadly conserved across genera including *Aedes*, *Culex*, and *Sabethes*, although species-specific deviations were evident. For example, RT elements in *Aedes albopictus* were largely confined to two chromosomes, whereas *Ochlerotatus sierrensis* showed enrichment toward both chromosomal ends and central regions.

Despite interspecies variation in absolute abundance and chromosomal preference, normalized profiles revealed a broadly conserved tendency toward hotspot formation across mosquito lineages. While these patterns indicate non-random clustering of EVEs, their functional and evolutionary significance remains to be elucidated.

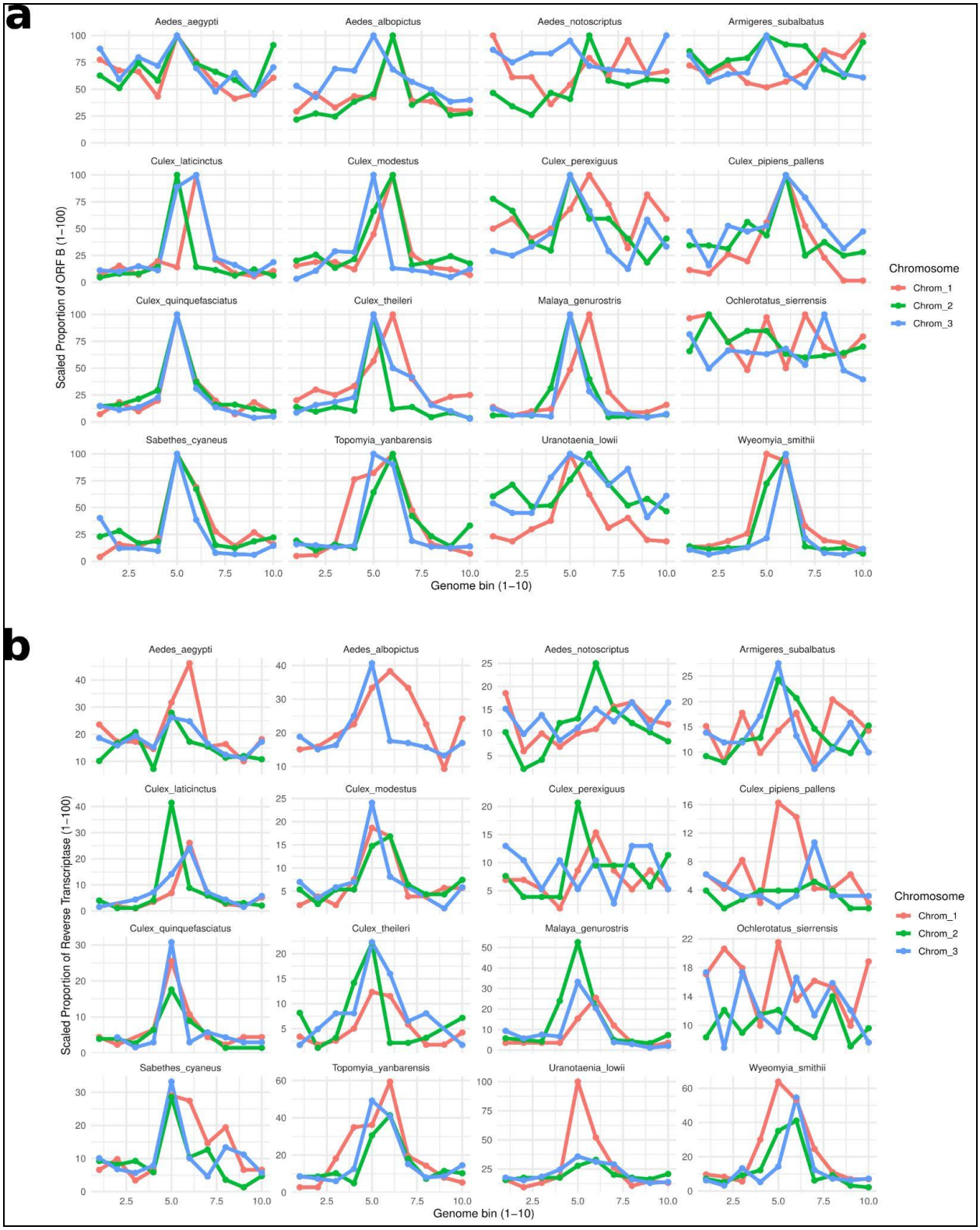


Fig. 5: Genome-wide distribution of ORF-B and Reverse Transcriptase homologs across mosquito genomes. (a) Distribution of ORF-B (*TED*-like) homologs across 16 mosquito species with chromosome-level genome assemblies. Chromosomes were segmented into 10 kb non-overlapping bins,

and ORF-B copy number was quantified per bin. Bin-wise counts were normalized by chromosome length to enable comparison across chromosomes and species. (**b)** Distribution of RT (Cladosporium-like) homologs across the same mosquito genomes, using the same binning and normalization approach as in (a).

Genomic Context of EVE Integration Sites

To investigate the genomic environments associated with EVE integration, we analyzed the 5-kb flanking regions surrounding *TED*-like and *Clado*–like EVEs within mosquito genomes. Protein domains in these loci were annotated using Pfam entries from the Conserved Domain Database (CDD) (Supplementary Table 4: Sheets 1-2).

***Trichoplusia ni TED*-like EVE Flanking Regions:** In the regions flanking *TED* virus ORF-B homologs, the most abundant domains were RT_RNaseH (pfam17917), RVT_1 (pfam00078), Integrase_H2C2 (pfam17921), and rve (pfam00665), which overlap with the ORF-B/pol region due to the inclusion of 5-kb extensions (Supplementary. Fig. 4a; Supplementary Table 4, Sheet 1). Additional enriched domains included those associated with TEs, such as reverse transcriptase, Gag, retrotransposon gag, DDE transposases, Gypsy elements, and retroviral proteases (Asp_protease_2, Peptidase_A17, RVP). Nucleic acid-binding and chromatin-associated domains (zinc-finger H2C2, PHD, fibrinogen-like proteins) were also frequently identified.

Functional enrichment analysis using Pfam-to-GO mapping identified DNA integration as the most significantly enriched biological function in these flanking regions (Supplementary. Fig. 4b). Other enriched categories included catalytic activity, RNA–DNA hybrid ribonuclease activity, nucleic acid binding, translation, ribosomal components, and aspartic-type endopeptidase activity.

***Cladosporium fulvum T-1–like* EVE Flanking Regions:** Analysis of the 5-kb flanking regions of *Cladosporium fulvum T-1 virus* RT homologs showed a similar enrichment pattern (Supplementary. Fig. 4c; Supplementary.Table 4, Sheet 2). The most abundant domains overlapped with the RT/polyprotein region, including RT_RNaseH, RVT_1, Integrase_H2C2, RVP_2 (pfam08284), and rve. Additional enriched domains included TE-associated motifs such as Gag, Gypsy, DDE transposases, retrotransposon gag, as well as protease domains (Asp_protease_2, Peptidase_A17). Other frequently detected motifs included zf-CCHC, SAP, PHD, histone-like proteins, and Tys_capsid.

GO enrichment analysis for these flanking regions also identified DNA integration as the dominant functional category (Supplementary. Fig. 4d), along with enrichment for catalytic activity, nucleic acid binding, RNA–DNA hybrid ribonuclease activity, zinc ion binding, and DNA binding. Overall, these analyses demonstrate that *Trichoplusia ni TED*-like and *Cladosporium fulvum T-1*–like EVEs are consistently integrated within TE-rich, retroelement-associated genomic regions in mosquito genomes. The strong enrichment of

retrotransposon-related domains and integration functions in both sets of flanking regions was supported by both domain and GO enrichment (Supplementary. Fig. 3a-b, c-d; S.Table 4, Sheets 1–4).

Syntenic organization of *TED*-like and *Clado*-like EVEs

To investigate whether the identified EVEs persist as complete viral integrations or as fragmented genomic remnants, we examined the genomic organization of *TED*-like and *Clado*-like loci across nineteen mosquito genomes. Using the same dataset generated for the analysis of flanking genomic environments, we analyzed genomic regions spanning ±5 kb around each Pol-associated EVE. The presence, co-occurrence, and relative arrangement of open reading frame (ORF)-associated protein domains were assessed to evaluate the extent of viral genome preservation and to determine whether ancestral viral genomic architectures remain detectable within mosquito genomes.

The canonical *TED* retrovirus of *Trichoplusia in TED* comprises three ORFs: ORF-A (gag), ORF-B (pol), and ORF-C (env). Among *TED*-like EVEs, ORF-B-associated protein domains were the most prevalent and were detected in all nineteen mosquito genomes. The Pol region was characterized by a conserved arrangement of reverse transcriptase (Pfam00078), retroviral aspartyl protease (Pfam00665), RNase H (Pfam17917), and integrase zinc-binding (Pfam17921) domains, indicating preservation of the ancestral Pol architecture across multiple loci. ORF-A-associated domains (Pfam03732) co-occurred with Pol-associated domains in seventeen mosquito genomes (Supplementary Table 5). Canonical Env-associated domains were not detected; however, ORF-B-associated domains co-occurred with a Baculo_F domain (Pfam12259)[54], encoding a baculovirus F-like fusion protein proposed to function as an evolutionary analogue of the retroviral Env protein in errantiviruses, in seventeen mosquito genomes (Supplementary Table 5). As a representative example, synteny analysis of the Aedes aegypti genome identified 45 loci containing both ORF-A- and ORF-B-associated domains (Supplementary Fig. 5a) and seven loci containing ORF-B-associated domains together with Baculo_F domains (Supplementary Fig. 5b). These two domain combinations occurred in distinct loci, and no integration simultaneously retained ORF-A-, ORF-B-, and Baculo_F-associated domains, indicating that complete *TED* genomic organization has not been preserved in mosquito genomes.

A similar pattern was observed for *Cladosporium fulvum T-1*-like elements. The predicted Pol region was consistently represented by the conserved suite of reverse transcriptase (Pfam00078), retroviral aspartyl protease (Pfam00665), RNase H (Pfam17917), and integrase zinc-binding (Pfam17921) domains. Moreover, these domains generally occurred in a conserved order across loci, reflecting preservation of ancestral retroelement microsynteny. Across the nineteen mosquito genomes, ORF-A-associated domains co-occurred with Pol-associated domains in all species examined, with locus counts ranging from 4 to 120 per genome (Supplementary Table 5).

In the representative Aedes aegypti synteny analysis, 70 loci contained both ORF-A- and ORF-B-associated domains (Supplementary Fig. 6). No loci retained additional viral ORFs consistent with a complete ancestral genome organization. However, Baculo_F-associated domains, suggestive of ORF-C-derived sequences, were detected in eight genomes (*Uranotaenia lowii*, *Culex quinquefasciatus*, *Aedes koreicus*, *Culex pipiens pallens*, *Aedes albopictus*, *Ochlerotatus camptorhynchus*, *Culex modestus*, and *Culex theileri*).

Collectively, these analyses indicate that *TED*-like and *Clado*-like EVEs are predominantly retained as fragmented viral elements rather than complete proviral integrations. Although the Pol region consistently preserved its ancestral domain organization, the remaining viral ORFs were only partially retained. *TED*-like loci contained either ORF-A or Baculo_F-associated domains adjacent to Pol-associated regions, but never all three components within a single locus. Similarly, *Clado*-like EVEs preserved only partial ancestral genome organization. Nevertheless, the widespread conservation of Pol-domain architecture and microsynteny across diverse mosquito lineages demonstrates that these elements retain recognizable signatures of their ancestral viral genomes despite extensive structural fragmentation.

Distribution of Defense Protein Domains in ±5 kb Flanking Regions of *TED* and *Clado*-like EVEs

To determine whether identified EVEs are associated with genomic regions enriched in for host defense-related proteins, the ±5 kb regions flanking *TED*- and *Clado*-like EVEs were analyzed for conserved PIWI-associated (pfam02170|PAZ, pfam02171|PIWI, pfam16487|ArgoMid, pfam16486|ArgoN, pfam16488|ArgoL2, pfam08699|ArgoL1) and prokaryotic defense-related protein domains using RPS-BLAST against Pfam profiles (Supplementary Table 6). Canonical PIWI-associated domains were detected near both *TED* and *Clado*-like EVEs, but only in a limited number of genomic loci (Supplementary Fig 7 a-b). In TED-associated regions, three to six PIWI-related domains (pfam02170|PAZ, pfam02171|PIWI, pfam16487|ArgoMid) were identified in four of the 19 mosquito species examined: *Aedes koreicus, Aedes japonicus, Ochlerotatus sierrensis,* and *Armigeres subalbatus* (Supplementary Fig 7a). Among these, the pfam02171|PIWI domain was most enriched, particularly in *Aedes koreicus* (19 copies), whereas PAZ- and ArgoMid-related domains were present at much lower frequencies. These findings suggest the presence of partial, rather than fully conserved, canonical PIWI-domain architectures near TED-derived EVEs.

Similarly, *Clado*-like EVEs were also associated with five out of six PIWI-related domains (pfam02170|PAZ, pfam02171|PIWI, pfam16486|ArgoN, pfam16488|ArgoL2, pfam08699|ArgoL1) across 19 mosquito genomes (Supplementary Table 6), although at lower overall abundance (Supplementary Fig 7b). The most frequently detected domains in these regions were pfam02171|PIWI and pfam02170|PAZ. Additional Argonaute-associated domains, including pfam16486|ArgoN, pfam16488|ArgoL2, and pfam08699|ArgoL1, were occasionally observed in *Ochlerotatus sierrensis*. Overall, the limited and fragmented distribution of

PIWI-associated domains within the ±5 kb flanking regions suggests that these EVEs are not consistently located adjacent to complete, conserved piRNA pathway genes. However, because the analysis was restricted to a 5 kb window on either side of each EVE, canonical PIWI proteins or complete piRNA pathway genes located farther from the integration sites may not be detected. Therefore, these results should not be interpreted as evidence for the absence of PIWI genes in the surrounding genomic landscape.

In addition to PIWI-associated proteins, numerous prokaryotic defense-related domains were identified in the flanking regions of both EVE groups (Supplementary Fig 7. c-d). *TED*-associated EVEs exhibited a broader and more abundant repertoire of defense domains compared to *Clado*-like EVEs. The most prevalent domains in TED-flanking regions included reverse transcriptase-related domains (RVT_1), trypsin-like peptidases, protein kinase domains, DEAD-box helicases, ankyrin-repeat proteins, helicase-associated modules, AAA+ ATPases, and restriction-associated nuclease-related proteins. Several domains characteristic of bacterial antiviral defense systems, including abortive infection-associated proteins, reverse transcriptase-related domains, TIR-like domains, viral helicases, and nuclease-associated proteins, were also detected.

Clado-like EVEs also showed enrichment of defense-related domains, but with lower diversity and abundance (Supplementary Fig. 7d). Reverse transcriptase-associated domains were the most prominent feature across most mosquito genomes, with sporadic occurrences of AAA+ ATPases, helicases, TIR-associated proteins, and nuclease-related domains. Recurrent associations between EVEs and defense-related domains were particularly evident in several mosquito species, including *Aedes aegypti, Aedes albopictus, Topomyia yanbarensis,* and *Aedes koreicus*.

Overall, these results indicate that mosquito EVEs are frequently located in genomic regions containing diverse defense-related protein domains, including limited PIWI-associated domains and a broader repertoire of domains characteristic of prokaryotic antiviral defense systems. Although canonical PIWI-domain architectures were only sporadically detected within the ±5 kb flanking regions, the recurrent association of EVEs with multiple defense-related domains suggests that these genomic neighborhoods may have been shaped by long-term evolutionary interactions between mosquitoes and viruses. Because the analysis was restricted to a 5 kb window surrounding each EVE, additional defense genes located farther from the integration sites cannot be excluded.

Transcriptomic Evidence for EVE Activity in *Culicinae* Mosquitoes

To assess the functional potential and activity of EVEs, we analyzed 28 publicly available assembled *Culicinae* transcriptomes using *EEFinder*. This approach identified 1,448 EVE-related sequences in 12 of 28 transcriptomes, indicating that viral elements are prevalent but exhibit uneven expression across samples (Supplementary Table 2, Sheets1-2). Low detection rates in certain transcriptomes may result from factors such as tissue type, developmental stage, sequencing depth, assembly quality, or conditional EVE activity. The

absence of detectable EVE transcripts in some datasets likely reflects low, transient, or context-specific expression rather than true genomic absence.

The detected EVE-related sequences were classified by genome type as follows: dsDNA (1,309), dsDNA-RT (2), ssDNA(+) (9), ssRNA-RT (67), ssRNA(−) (45), ssRNA(+) (7), and unknown (9). Taxonomic classification assigned these elements to 44 viral families, 257 genera, and 440 unique viral species, demonstrating substantial viral diversity within Culicinae transcriptomes (Supplementary Fig. 8 a-d). Analysis of the 20 most prevalent multiple viral species distribution patterns similar to those observed in the genomic EVE analysis. The most frequently detected EVE transcripts corresponded to *TED*-like *virus* (52.1%) and *Clado*-like *virus* (10.3%), alongside EVEs related to *Fadolivirus algeromassiliense, Bracoviriform inaniti,* and *Tupanvirus deep* ocean. These results indicate that certain EVEs remain transcriptionally active and their presence in transcriptomes broadly reflects the genomic EVE landscape.

PCR and sequence based validation for presence of EVEs in *Aedes Aegypti*

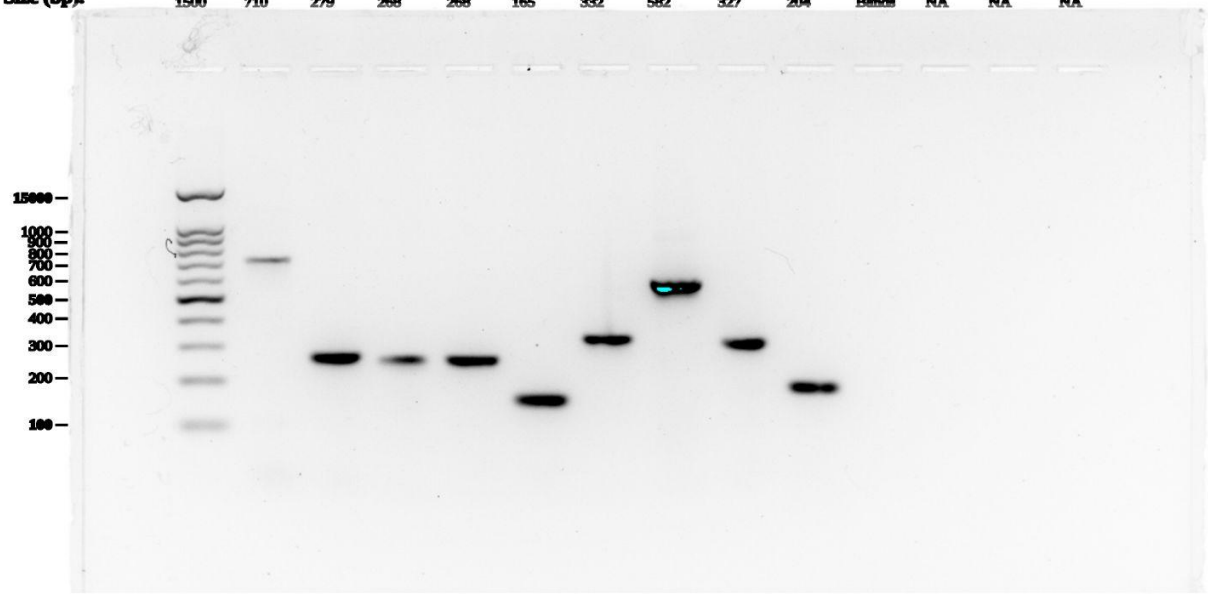
To experimentally validate computationally predicted endogenous viral elements (EVEs), representative TED-like and Clado-like loci were amplified from Indian Aedes aegypti genomic DNA using locus-specific primers. PCR products were resolved on 2% (w/v) agarose gels and visualized. A mitochondrial COX1 amplicon served as a positive control for DNA quality and PCR performance, whereas a reagent blank was included as a negative control in all experiments 55. Gel images were acquired using the iBright™ Imaging System (Thermo Fisher Scientific) and processed in Fiji (v2.0.2) by grayscale conversion, inversion, and cropping without altering brightness or contrast. Final figures were assembled using Inkscape (v1.4.4) (Fig. 6a,b).

PCR amplification generated products of the expected sizes for all selected TED-like and Clado-like EVE loci, while the COX1 positive control amplified successfully and no amplification was detected in the reagent blank, confirming the specificity of the assay (Fig. 6a,b) Selected PCR amplicons were validated by bidirectional Sanger sequencing, and consensus and individual sequences assembled from the chromatogram were translated and queried against the NCBI non-redundant protein database using BLASTX to identify homologous viral proteins.

a

PCR Validation of TED-like Endogenous Viral Elements in Mosquito Genomic DNA

Lane No.	M	2	3	4	5	6	7	8	9	10	11	12	13	14
Lane Label	DNA Ladder	COX1 Positive Control	1.1	1.4	1.6	1.7	1.8	1.10	1.11	1.13	Reagent Blank	Empty	Empty	Empty
Expected Size (bp).	100 - 1500	710	279	268	268	165	332	582	327	204	No Bands	NA	NA	NA

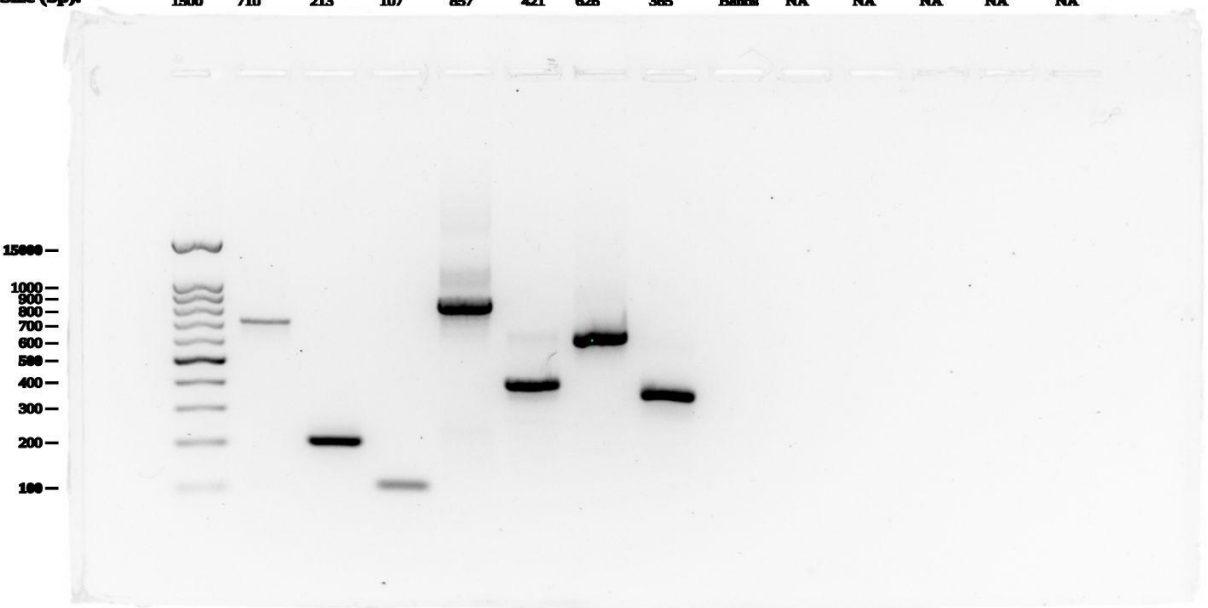


Gel Concentration: 2% (w/v)Agarose Gel

b

PCR Validation of Clado-like Endogenous Viral Elements in Mosquito Genomic DNA

Lane No.	M	2	3	4	5	6	7	8	9	10	11	12	13	14
Lane Label	DNA Ladder	COX1 Positive Control	2.3	2.4	2.6	2.7	2.8	2.9	Reagent Blank	Empty	Empty	Empty	Empty	Empty
Expected Size (bp).	100 - 1500	710	213	107	857	421	626	365	No Bands	NA	NA	NA	NA	NA



Gel Concentration: 2% (w/v)Agarose Gel

Fig. 6: Experimental validation of predicted TED-like and Clado-like EVEs in Aedes aegypti. PCR amplicons were resolved on a 2% (w/v) agarose gel. Lane M, DNA ladder (100-1500 bp); lane 2, COX1 positive control (710 bp); (a) lanes 3-8, predicted *TED*-like EVE loci with expected amplicon sizes of 279, 268, 268, 165, 332, 582, 327, and 204 bp, respectively; lane 11, reagent blank (no amplification); lanes 12-14, empty lanes. (b) lanes 3-8, predicted *Clado*-like EVE loci with expected amplicon sizes of 213, 107, 857, 421, 626, and 365 bp, respectively; lane 9, reagent blank (no amplification); lanes 10-14, empty lanes.

PCR amplification generated products of the expected sizes for *TED*-like and *Clado*-like loci, while the COX1 positive control amplified successfully and no amplification was observed in the reagent blank. PCR amplification generated products of the expected sizes for *TED*-like and *Clado*-like loci, while the COX1 positive control amplified successfully and no amplification was observed in the reagent blank.

Sanger sequencing validation of PCR-amplified EVE loci

Bidirectional Sanger sequencing was performed for all PCR amplicons, including the COX1 positive control. High-quality forward sequences were obtained for 11 of the 13 EVE loci, whereas locus 1.4 failed to yield a usable forward read. Forward sequences from TED-like loci 1.1, 1.6, 1.8, 1.11, and 1.13 recovered top BLASTX matches to Trichoplusia ni TED virus ORF B, supporting the predicted EVE. Similarly, Clado-like loci 2.6, 2.8, and 2.9 matched the reverse transcriptase of Cladosporium fulvum T-1 virus. In contrast, locus 2.7 recovered its highest-scoring match to Atlantic salmon swim bladder sarcoma virus, while loci 1.7, 1.10, 2.3, and 2.4 did not produce significant viral BLASTX matches. As expected, the COX1 positive control matched mosquito cytochrome oxidase subunit I (Table S8) .High-quality reverse sequences were obtained for 13 of the 13 EVE loci. Reverse BLASTX analysis largely corroborated the forward-read results. TED-like loci 1.1, 1.6, 1.8, 1.11,and 1.13 again recovered Trichoplusia ni TED virus ORF B as the top viral match. Locus 1.7 instead matched the RNA-dependent RNA polymerase of Aedes aegypti To virus 1. Among the Clado-like candidates, locus 2.8 retained a top match to Cladosporium fulvum T-1 virus, whereas loci 2.6, 2.7, and 2.9 recovered highest-scoring matches to unrelated retroviral proteins rather than the expected Clado-like lineage. The COX1 reverse sequence again recovered the expected mosquito cytochrome oxidase subunit I protein.

Manual inspection of forward and reverse chromatograms identified concordant sequence regions for 12 amplicons, comprising the COX1 positive control and loci 1.1, 1.6, 1.8, 1.10, 1.11, 1.13, 2.3, 2.6, 2.7, 2.8, and 2.9. No consensus sequence could be generated for locus 1.4 because no high-quality Sanger sequence was recovered, for locus 2.4 because no concordant sequence was identified following manual alignment inspection. BLASTX analysis of the concordant consensus sequences confirmed the predicted viral lineage for four TED-like loci (1.1, 1.8, 1.11, and 1.13) and three Clado-like loci (2.6, 2.8, and 2.9). Consensus sequences from loci 1.7, 1.10

and 2.3 produced no significant viral protein matches despite successful sequence assembly. Locus 1.6 recovered a weak match to Baboon endogenous virus strain M7 rather than TED virus, while locus 2.7 remained associated with a retroviral protein outside the expected Clado-like lineage. The assembled COX1 consensus sequence recovered mosquito cytochrome c oxidase subunit I, confirming successful amplification and sequencing of the positive control.

Overall, concordant Sanger sequence analysis validated 7 of the 10 evaluable EVE candidates, including four TED-like (1.1, 1.8, 1.11, and 1.13) and three Clado-like (2.6, 2.8, and 2.9) loci. Three loci (1.6, 1.10, and 2.3) did not retain the expected viral signatures following consensus analysis, while locus 2.7 consistently recovered an alternative retroviral lineage. No consensus sequence could be generated for loci 1.4 and 2.4. This sequence-level validation supports the authenticity of the majority of experimentally amplified EVE candidates while highlighting a small number of loci with ambiguous or inconsistent viral affiliations requiring cautious interpretation.

Discussion:

EVEs are genetic remnants of ancient viral infections embedded within mosquito genomes, preserving a molecular record of historical virus–host interactions [15,16]. Analysis of 19 Culicinae RefSeq mosquito genomes identified a diverse array of EVEs originating from both known and previously uncharacterized viral lineages. These findings suggest that viral sequence integration has occurred frequently and persistently throughout mosquito evolution [15,17]. The presence of EVEs in all the examined species indicates their evolutionary importance and reveals a complex mosaic of virus-derived sequences in modern mosquito genomes. Rather than passive remnants, EVEs act as molecular fossils, retaining traces of viral lineages that may no longer be present [16]. These observations provide evidence that viral integrations contribute to genetic diversity and evolutionary processes in mosquitoes [17].

Notably, retrovirus-like elements, particularly those related to *Metaviridae* and *Errantivirus*, predominated across the analyzed genomes. This pattern, which is also observed in other insects, suggests that gypsy-like LTR retroelements have contributed to repetitive DNA content and have persisted over extended evolutionary periods [56,57]. Their widespread distribution across *Culicinae* genomes implies that many integrations predate major speciation events. Furthermore, the presence of similar EVE repertoires in Indian and globally derived *Aedes aegypti* genomes further indicates that many viral integrations are evolutionarily conserved across geographically distinct populations consistent with an ancient origin and long-term persistence. Together, these observations support a model in which ancient integrations have been stably maintained over evolutionary time. The frequent association of EVEs with retrotransposon-derived sequences supports the hypothesis that endogenous retrotransposons may facilitate viral cDNA formation

and integration, contributing to the high abundance and diversification of EVEs in mosquito genomes [26,33,58].

Beyond retroviral sequences, we identified numerous EVEs derived from non-retroviral viral groups, including *Rhabdoviridae, Chuviridae, Flaviviridae, Reoviridae*, and members of *Bunyavirales*. Many of these groups encompass arboviruses and insect-specific viruses, which are major contributors to mosquito EVE content [30,33]. Consistent with previous studies, *Aedes* species harbored the highest EVE loads, supporting the idea that transposable elements–rich genomes facilitate the accumulation and retention of viral integrations [11,26,33] Previous studies have shown that EVEs can generate piRNAs that suppress cognate viral infections, suggesting a role in heritable antiviral memory [10,26,27,59]. Although functional validation was not performed, the diversity and distribution of EVEs support the idea that at least a subset of the discovered EVEs has potential involvement in antiviral responses and long-term host–virus coevolution.

Our analyses also revealed a non-random chromosomal distribution of EVEs, with several regions exhibiting elevated densities of viral integrations, forming distinct hotspots. Similar clustering patterns have been reported for transposable elements and endogenous viral sequences in other eukaryotic genomes and may reflect local chromatin environments, transposable element activity, or long-term selective retention [18,60]. Similar clustering patterns have been reported in *Aedes* mosquitoes, where EVEs are enriched in repeat-rich genomic regions and piRNA clusters, as well as in plant genomes containing endogenous geminivirus-like elements (EGVs), suggesting that hotspot formation may represent a common feature of endogenous viral sequence organization across diverse eukaryotic lineages [11,19]. These patterns support the idea that viral integration is influenced by genome architecture rather than occurring randomly. Whether these hotspots contribute to antiviral defense or genome regulation remains to be determined.

Unexpected*ly, we detected sequences showing similarity to members of the Nucleocytoviricota, including Mimiviridae, Marseilleviridae, Poxviridae, Iridoviridae, Phycodnaviridae, and Pandoraviridae*-like homologs. Such signals are rarely reported in mosquito genomes [61]. However, members of *Nucleocytoviricota*, including *Marseillevirus*-like viruses, have recently been detected in Aedes aegypti populations, with evidence of persistence across developmental stages and potential transmission pathways, highlighting that giant virus–mosquito associations may occur in natural settings. Nevertheless, given the reliance on homology-based inference, these findings may reflect conserved domains, assembly artifacts, or potential contamination [62].

TED and Clado-like virus elements were among the most abundant EVEs identified in both global and Indian *Aedes aegypti* datasets. This observation parallels recent virome-based studies in sylvatic mosquito populations[63], though our findings are distinct in demonstrating these elements as integrated, rather than transient, viral signatures. These elements share features with gypsy/errantivirus-like LTR retroelements, including conserved gag and pol domains. In addition, a subset of *TED*-like EVEs retained Baculo_F domains adjacent to *pol*-associated regions, consistent with the alternative envelope architecture of insect errantiviruses, although no integration preserved all three canonical genomic components [54,64]. Their conservation across

genomes suggests an ancient origin and long-term persistence, while phylogenetic and clustering analyses revealed several divergent lineages that may represent previously undescribed TED-like retroelements, highlighting the extensive unexplored diversity of retrovirus-like sequences in mosquito genomes.

Analysis of genomic regions ± 5kb flanking *TED*-like and *Clado*-like EVEs revealed enrichment of retroelement-associated domains and a limited number of Piwi/Argonaute-related domains. Previous studies have shown that mosquito EVEs are frequently enriched within piRNA clusters and contribute to the production of antiviral piRNAs. Therefore, the genomic context observed here is consistent with a potential interaction between EVEs and small RNA-mediated antiviral pathways. However, the relatively low frequency of piRNA-associated domains detected in our analyses suggests that only a subset of EVEs may be directly linked to antiviral functions. Alternatively, the observed pattern may reflect the limited genomic window analyzed or the complexity of piRNA cluster organization in mosquito genomes[33].

Analysis of EVE-flanking regions revealed a limited number of Pfam domains also found in proteins associated with prokaryotic defense systems, such as retron- and abortive infection-related proteins. Because these domains are widely distributed and occur in various protein families, their presence does not indicate intact prokaryotic defense systems in mosquito genomes. Instead, this likely reflects conserved functional domains or deep evolutionary relationships. The association of these domains with EVEs raises important questions about their evolutionary origins and potential functional roles that merit further investigation.

A key strength of this study is the integration of genome-scale computational analysis with experimental validation. Selected EVEs were confirmed in Indian *Aedes aegypti* populations by PCR, and transcriptomic data supported the active transcription of some integrations, underscoring their authenticity. Several limitations remain: homology-based identification may not distinguish ancient viral integrations from highly diverged mobile elements, reference genomes may not capture full population-level diversity, and functional roles of EVEs require further characterization.

While this study offers important insights, several limitations should be noted. EVE identification based on homology may not distinguish ancient integrations from highly diverged mobile elements. Reference genomes also do not capture full population-level variation. Moreover, functional roles remain largely uncharacterized despite experimental validation of presence and transcription. Future work integrating long-read sequencing, population genomics, and functional assays will be essential. Beyond their biological roles, EVEs provide a valuable paleovirological record of historical virus–host interactions and ancient viral diversity, offering insights into viral lineages that may no longer circulate in contemporary populations [15,25].

Overall, our results reveal a highly diverse and dynamic landscape of EVEs in *Culicinae* mosquitoes. The coexistence of ancient conserved integrations, lineage-specific EVEs, and previously uncharacterized retrovirus-like elements underscores the long-term and ongoing impact of viral endogenization on mosquito genome evolution. By expanding the known

repertoire of mosquito EVEs and identifying putative functional associations with genome organization and antiviral pathways, this study provides a foundation for future work on the evolutionary and biological consequences of viral integration in mosquito vectors.

Conclusion

This study provides the most comprehensive comparative characterization of EVEs in *Culicinae* mosquitoes to date by integrating genome-wide analyses, transcriptomic evidence, and experimental validation. Across 19 mosquito genomes, we identified 77,141 putative EVEs representing 1,538 viral species, with *TED*-like and *Clado*-like retroviral elements accounting for approximately 62% of all identified EVEs. These elements exhibit non-random genomic distributions, strong associations with transposable element-rich regions, and extensive fragmentation, consistent with ancient integration and long-term evolutionary persistence. Phylogenetic analyses uncovered multiple previously unrecognized EVE lineages, expanding the known diversity of mosquito-associated viruses and suggesting recurrent viral integration throughout mosquito evolution. PCR validation confirmed the genomic integration of representative *TED* and *Clado* EVEs in Indian *Aedes aegypti*, while transcriptomic analyses demonstrated that a subset of EVEs is transcriptionally active. Together, these findings establish a valuable genomic resource for investigating the evolutionary history of virus–host interactions and provide a foundation for future studies exploring the roles of EVEs in mosquito genome evolution, antiviral immunity, and vector biology.

Data Availability

All mosquito genomes analyzed in this study are publicly available from the NCBI FTP server under the accession numbers listed in Supplementary Table S1. The identified EVEs, along with custom scripts for data analysis and visualization, are available on GitHub at https://github.com/sharmavikas-niv/Mosquito-paleovirology.git.

Acknowledgement

This work utilized the High Performance Computing Facility "NAKSHATRA," established under the Indian Council of Medical Research PM-Ayushman Bharat Health Infrastructure Mission (PM-ABHIM) project (Grant No. VIR/35/2022/ECD-1), to perform the bioinformatics analyses.

Authors Contributions

VS and NK conceived and designed the study. NK provided overall scientific guidance and supervision throughout the study. VS coordinated the project, performed data collection, conducted bioinformatics analyses with NV, interpreted the results, and managed the overall

investigation. AFA and RMJ implemented the EEfinder pipeline using the Nakshatra high-performance computing (HPC) facility. AKS downloaded and curated transcriptome (TSA) datasets. NV designed the primers, and NV and AKS optimized and validated them for experimental analyses. SK provided mosquito DNA samples, while AKS performed PCR amplification and Sanger sequencing for validation of the predicted endogenous viral elements (EVEs), assembled and curated the Sanger sequencing data, and conducted sequence homology analyses. HA analyzed EVEs associated with human disease. VKM, HK, and NK contributed to data interpretation and refinement of the scientific content. VS and NV wrote the original draft of the manuscript. NK, VKM, and HK critically reviewed the manuscript and provided intellectual input. VS, HK, AP, and VKM revised and edited the manuscript. All authors reviewed and approved the final manuscript.

Declaration of Generative AI and AI-Assisted Technologies in Manuscript Preparation

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) and Grammarly solely to improve the text's clarity, grammar, and readability.

References

1. Arensburger, P. *et al.* Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science* **330**, 86 (2010).
2. Neafsey, D. E. *et al.* Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* **347**, 1258522 (2015).
3. Matthews, B. J. *et al.* Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature* **563**, 501–507 (2018).
4. Melo, E. S. de & Wallau, G. L. Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. *PLOS Genet.* **16**, e1008946 (2020).
5. Franklinos, L. H. V., Jones, K. E., Redding, D. W. & Abubakar, I. The effect of global change on mosquito-borne disease. *Lancet Infect. Dis.* **19**, e302–e312 (2019).
6. Weaver, S. C. & Reisen, W. K. Present and future arboviral threats. *Antiviral Res.* **85**, 328–345 (2010).

7. Nelson, A. N. & Ploss, A. Emerging mosquito-borne flaviviruses. *mBio* **15**, e02946-24.
8. Prince, B. C., Walsh, E., Torres, T. Z. B. & Rückert, C. Recognition of Arboviruses by the Mosquito Immune System. *Biomolecules* **13**, 1159 (2023).
9. Campbell, C. L., Black, W. C., Hess, A. M. & Foy, B. D. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* **9**, 425 (2008).
10. Tassetto, M. *et al.* Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. *eLife* **8**, e41244 (2019).
11. Palatini, U., Contreras, C. A., Gasmi, L. & Bonizzoni, M. Endogenous viral elements in mosquito genomes: current knowledge and outstanding questions. *Curr. Opin. Insect Sci.* **49**, 22–30 (2022).
12. Crochu, S. *et al.* Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. *J. Gen. Virol.* **85**, 1971–1980 (2004).
13. Roiz, D., Vázquez, A., Seco, M. P. S., Tenorio, A. & Rizzoli, A. Detection of novel insect flavivirus sequences integrated in Aedes albopictus (Diptera: Culicidae) in Northern Italy. *Virol. J.* **6**, 93 (2009).
14. Lequime, S. & Lambrechts, L. Discovery of flavivirus-derived endogenous viral elements in Anopheles mosquito genomes supports the existence of Anopheles-associated insect-specific flaviviruses. *Virus Evol.* **3**, vew035 (2017).
15. Katzourakis, A. & Gifford, R. J. Endogenous Viral Elements in Animal Genomes. *PLOS Genet.* **6**, e1001191 (2010).
16. Holmes, E. C. The Evolution of Endogenous Viral Elements. *Cell Host Microbe* **10**, 368–377 (2011).

17. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet. 2012 134* **13**, 283–296 (2012).
18. Gilbert, C. & Feschotte, C. Endogenous viral elements: evolution and impact. *Virologie* **20**, 158–173 (2016).
19. Sharma, V. *et al.* Large-scale survey reveals pervasiveness and potential function of endogenous geminiviral sequences in plants. *Virus Evol.* **6**, 71 (2020).
20. Diop, S. I. *et al.* Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Sci. Rep.* **8**, 1–9 (2018).
21. Aswad, A. & Katzourakis, A. Paleovirology: The Study of Endogenous Viral Elements. in *Virus Evolution: Current Research and Future Directions* 273–292 (Caister Academic Press, 2016). doi:10.21775/9781910190234.10.
22. Patel, M. R., Emerman, M. & Malik, H. S. Paleovirology—ghosts and gifts of viruses past. *Curr. Opin. Virol.* **1**, 304–309 (2011).
23. Geuking, M. B. *et al.* Recombination of Retrotransposon and Exogenous RNA Virus Results in Nonretroviral cDNA Integration. *Science* **323**, 393–396 (2009).
24. Horie, M. *et al.* Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
25. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479–480**, 26–37 (2015).
26. Crava, C. M. *et al.* Population genomics in the arboviral vector Aedes aegypti reveals the genomic architecture and evolution of endogenous viral elements. *Mol. Ecol.* **30**, 1594–1611 (2021).

27. Palatini, U. *et al.* Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics* **18**, 512 (2017).
28. Dezordi, F. Z., Vasconcelos, C. R. dos S., Rezende, A. M. & Wallau, G. L. In and Outs of Chuviridae Endogenous Viral Elements: Origin of a Potentially New Retrovirus and Signature of Ancient and Ongoing Arms Race in Mosquito Genomes. *Front. Genet.* **11**, (2020).
29. Hofer, U. Antiviral ping-pong in mosquitoes. *Nat. Rev. Microbiol.* **18**, 542–543 (2020).
30. Blitvich, B. J. & Firth, A. E. Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. *Viruses* **7**, 1927–1959 (2015).
31. Guzman, H. *et al.* Characterization of Three New Insect-Specific Flaviviruses: Their Relationship to the Mosquito-Borne Flavivirus Pathogens. *Am. J. Trop. Med. Hyg.* **98**, 410–419 (2018).
32. Roundy, C. M., Azar, S. R., Rossi, S. L., Weaver, S. C. & Vasilakis, N. Insect-Specific Viruses: A Historical Overview and Recent Developments. *Adv. Virus Res.* **98**, 119–146 (2017).
33. Whitfield, Z. J. *et al.* The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the Aedes aegypti Genome. *Curr. Biol.* **27**, 3511-3519.e7 (2017).
34. Dias, Y. J. M., Dezordi, F. Z. & Wallau, G. da L. EEfinder, a general purpose tool for identification of bacterial and viral endogenized elements in eukaryotic genomes. *Comput. Struct. Biotechnol. J.* **0**, (2024).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

36. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

37. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**, e0163962 (2016).

38. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

39. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

40. Katoh, K. & Standley, D. M. MAFFT: Iterative Refinement and Additional Methods. *Methods Mol. Biol.* **1079**, 131–146 (2014).

41. Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics* **8**, 312 (2007).

42. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

43. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

44. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

45. van der Maaten, L. & Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

46. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

47. Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).

48. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).

49. Miesen, P., Ivens, A., Buck, A. H. & Rij, R. P. van. Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs. *PLoS Negl. Trop. Dis.* **10**, e0004452 (2016).

50. The Crystal Structure of Human Argonaute2 | Science. https://www.science.org/doi/10.1126/science.1221551.

51. Systematic and quantitative view of the antiviral arsenal of prokaryotes | Nature Communications. https://www.nature.com/articles/s41467-022-30269-9.

52. Okonechnikov, K., Golosova, O., Fursov, M., & the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).

53. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).

54. Pearson, M. N. & Rohrmann, G. F. Envelope gene capture and insect retrovirus evolution: the relationship between errantivirus and baculovirus envelope proteins. *Virus Res.* **118**, 7–15 (2006).

55. Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299 (1994).

56. Peccoud, J., Loiseau, V., Cordaux, R. & Gilbert, C. Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci.* **114**, 4721–4726 (2017).
57. Terzian, C., Ferraz, C., Demaille, J. & Bucheton, A. Evolution of the Gypsy endogenous retrovirus in the Drosophila melanogaster subgroup. *Mol. Biol. Evol.* **17**, 908–914 (2000).
58. Goic, B. *et al.* Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat. Commun.* **7**, 12410 (2016).
59. Suzuki, Y. *et al.* Non-retroviral Endogenous Viral Element Limits Cognate Virus Replication in Aedes aegypti Ovaries. *Curr. Biol. CB* **30**, 3495-3506.e6 (2020).
60. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
61. Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **46**, D708–D717 (2018).
62. Karalyan, Z. *et al.* Detection and infection dynamics of Marseillevirus-like viruses in *Aedes aegypti*: Evidence for vertical and human exposure pathways. *Acta Trop.* **280**, 108185 (2026).
63. Gómez-Palacio, A. *et al.* Metagenomic profiling of the insect-specific virome in non-urban mosquitoes (Culicidae: Culicinae) from Colombia's Northern inter-Andean valleys. *PLOS One* **20**, e0331552 (2025).
64. Malik, H. S. & Henikoff, S. Positive Selection of Iris, a Retroviral Envelope–Derived Host Gene in Drosophila melanogaster. *PLoS Genet.* **1**, e44 (2005).

Supplementary Files

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

- SupplementaryMosquitopaleovirologyVS.pdf