frontiers in Genetics

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fgene.2020.542437

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

*Filipe Zimmer Dezordi[1], Crhisllane Rafaele dos Santos Vasconcelos[2], Antonio Mauro Rezende[2] and Gabriel Luz Wallau[1]**

[1] Departamento de Entomologia, Instituto Aggeu Magalhães (IAM), Fundação Oswaldo Cruz (FIOCRUZ), Recife, Brazil,
[2] Departamento de Microbiologia, Instituto Aggeu Magalhães (IAM), Fundação Oswaldo Cruz (FIOCRUZ), Recife, Brazil

OPEN ACCESS

Edited by:
Horacio Naveira,
University of A Coruña, Spain

Reviewed by:
Jean-Michel Drezen,
Université de Tours, France
Carol D. Blair,
Colorado State University,
United States

***Correspondence:**
Gabriel Luz Wallau
gabriel.wallau@cpqam.fiocruz.br;
gabriel.wallau@fiocruz.br

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 12 March 2020
Accepted: 28 September 2020
Published: 22 October 2020

Citation:
Dezordi FZ, Vasconcelos CRdS, Rezende AM and Wallau GL (2020) 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. 11:542437. doi: 10.3389/fgene.2020.542437

Background: Endogenous viral elements (EVEs) are sequences of viral origin integrated into the host genome. EVEs have been characterized in various insect genomes, including mosquitoes. A large EVE content has been found in *Aedes aegypti* and *Aedes albopictus* genomes among which a recently described *Chuviridae* viral family is of particular interest, owing to the abundance of EVEs derived from it, the discrepancy among the chuvirus endogenized gene regions and the frequent association with retrotransposons from the BEL-Pao superfamily. In order to better understand the endogenization process of chuviruses and the association between chuvirus glycoproteins and BEL-Pao retrotransposons, we performed a comparative genomics and evolutionary analysis of chuvirus-derived EVEs found in 37 mosquito genomes.

Results: We identified 428 EVEs belonging to the *Chuviridae* family confirming the wide discrepancy among the chuvirus genomic regions endogenized: 409 glycoproteins, 18 RNA-dependent RNA polymerases and one nucleoprotein region. Most of the glycoproteins (263 out of 409) are associated specifically with retroelements from the Pao family. Focusing only on well-assembled Pao retroelement copies, we estimated that 263 out of 379 Pao elements are associated with chuvirus-derived glycoproteins. Seventy-three potentially active Pao copies were found to contain glycoproteins into their LTR boundaries. Thirteen out of these were classified as complete and likely autonomous copies, with a full LTR structure and protein domains. We also found 116 Pao copies with no trace of glycoproteins and 37 solo glycoproteins. All potential autonomous Pao copies, contained highly similar LTRs, suggesting a recent/current activity of these elements in the mosquito genomes.

Conclusion: Evolutionary analysis revealed that most of the glycoproteins found are likely derived from a single or few glycoprotein endogenization events associated with a recombination event with a Pao ancestral element. A potential functional Pao-chuvirus hybrid (named Anakin) emerged and the glycoprotein was further replicated through

retrotransposition. However, a number of solo glycoproteins, not associated with Pao elements, can be found in some mosquito genomes suggesting that these glycoproteins were likely domesticated by the host genome and may participate in an antiviral defense mechanism against both chuvirus and Anakin retrovirus.

Keywords: endogenous virus elements, mosquitoes, transposons, retrotransposons, chuvirus

INTRODUCTION

Viruses have long-term and intricate interactions parasitizing host cells and both viruses and hosts are subject to an endless arms race (Forterre and Prangishvili, 2009). A large body of evidence currently supports that virus/host interactions can occur at both the protein and nucleic acid levels. One clear example of the last, is that viral genomic sequences can be integrated into the host genome (Feschotte and Gilbert, 2012; Johnson, 2019). The process of viral genome integration is called endogenization and normally occurs as a "life cycle" stage in viral groups such as retroviruses and phage DNA viruses (Weiss, 2016). However, recent studies have shown that genomes, or genomic regions, of non-integrative viruses can also be found integrated into various eukaryotic genomes (Katzourakis and Gifford, 2010; Feschotte and Gilbert, 2012; Johnson, 2019). These viral loci have been called endogenous viral elements (EVEs).

Endogenization can occur by two main mechanisms: through non-homologous recombination mediated by double-strand break repair pathway of the host cell; or mediated by proteins, such as reverse transcriptases and integrases, from the endogenous retrotransposons—envelope-free retrovirus-like elements (Katzourakis and Gifford, 2010). Recent findings on mosquito genomes suggest that the latter mechanism is likely to be the most important, since the abundance and diversity of EVEs are positively correlated with retrotransposon abundance and activity (Palatini et al., 2017; Whitfield et al., 2017). Likewise, EVEs and some retrotransposons families were found in neighboring loci in the *Aedes aegypti* Aag2 genome assembly (Whitfield et al., 2017). Such association does not appear to be only a physical co-localization, but a result of putative antiviral mechanism mediated by the activity of some retrotransposon elements (Whitfield et al., 2017; Tasseto et al., 2019). Lastly, Crava et al. (2020) have shown that TEs from BEL-Pao superfamily are enriched into piRNA clusters found in the AaegL5 *Ae. aegypti* genome assembly and this enrichment can influence their association with EVEs.

EVEs, are generally found as fragments of exogenous viral genomes and it is therefore unlikely that they are able to generate new virus particles or infect new cells. Therefore, there are three non-mutually exclusive hypotheses regarding the fate and impact of these elements on the host genome: (i) EVEs may evolve neutrally accumulating mutations and degenerate over time; (ii) EVEs may be co-opted by the host genomes, giving rise to new functional host genes; and (iii) EVEs may play an antiviral role, generating small RNAs which degrade cognate exogenous viral RNA or non-functional viral proteins that hinders proper assembly/maturation of a new viral particles or blocks the viral receptor on the host cell surface (Robinson et al., 1981; Ito et al., 2013; Armezzani et al., 2014). The vast majority of studies on EVEs found in mosquito genomes focus on the role of piRNA production as a post-transcriptional regulatory mechanism of exogenous circulating viruses (Théron et al., 2013; Palatini et al., 2017; Whitfield et al., 2017). On the other side, there is only one study that considers the potential role of EVEs at the mRNA expression level (Pischedda et al., 2019).

Chuviridae is a recently discovered RNA viral family of negative-sense single-stranded viruses characterized by metatranscriptomic and bioinformatic analysis only (Shi et al., 2016). The information available on this family is limited to its distribution (it likely infects several arthropod species including mosquitoes), its variable genomic structure (unsegmented, bisegment and circular), and the presence of a number of EVEs in species from Amphipoda, Hemiptera, Coleoptera, Hymenoptera and Diptera genomes (Shi et al., 2016). An in-depth descriptive analysis of chuvirus-derived EVEs exists only for the *Ae. aegypti* Aag2 cell-line genome (Whitfield et al., 2017) displaying three intriguing features: a large abundance of chuvirus-derived EVEs compared to other viral families (it is the second most abundant EVE family, outnumbered only by the *Rhabdoviridae* family); an association with retroelements from the BEL-Pao superfamily; and a striking difference in the viral genome fragment endogenized—a much higher quantity of glycoprotein (Gly) compared to RNA dependent RNA polymerase (RdRp) and nucleoprotein (NP) sequences (Whitfield et al., 2017). Endogenization of different viral regions is expected to be influenced by the virus genome structure and orientation of the RNA replication process. Whitfield et al. (2017) have shown that EVEs derived from another non-segmented negative-sense single-stranded virus from the *Rhabdoviridae* family occur in the following order of abundance: NP − > Gly − > RdRp resembling the order in which the viral genomes is replicated in this viral family. On the other hand, *Chuviridae* viruses, which are similar to *Rhabdoviridae* in genomic structure (NP-Gly-RdRp), shows a very different endogenization pattern, with eighty-seven endogenous Gly sequences, only four from RdRp and no endogenized nucleoproteins (Whitfield et al., 2017). It may indicate that chuvirus-derived EVEs found in *Ae. aegypti* genome were either derived from an exogenous chuvirus with segmented or non-segmented genome and for some reason Gly has been majorly endogenized, or that the endogenization of chuvirus genomic regions occurred evenly at the beginning but only Gly was maintained through the evolutionary time.

Endogenous repetitive elements are abundant in metazoan genomes and mosquito genomes from the *Aedes* genus are particularly full of retrotransposons/retroviruses (Nene et al., 2006; Goubert et al., 2015; Whitfield et al., 2017). One of the most abundant LTR retrotransposons/retroviruses are from the

BEL-Pao superfamily which is also very abundant and widespread in other metazoan genomes (Malik et al., 2000). Elements from this group have two long terminal repeat (LTR) regions, between 100 and 900 base pairs, and two coding regions—a capsid protein with a GAG domain, and a polyprotein, which commonly has four domains: aspartic protease (PR), reverse transcriptase (RT), RNAse H (RH), and integrase (INT). Moreover, at least three elements (Roo, Tas, and Cer-7) have an envelope-like protein (Gly) downstream of the polyprotein, which were acquired from *Gypsy*, Phlebovirus, and Herpesvirus, respectively (Felder et al., 1994; Browning et al., 1996; Malik and Henikoff, 2005).

In view of all the intriguing aforementioned chuvirus/BEL-Pao/host features, we investigated in depth which biological phenomenon has generated the high abundance of chuvirus glycoproteins found in mosquito genomes and examined the role these glycoproteins may play in BEL-Pao retroelements and mosquito biology. Our results showed, for the first time, that these EVEs are broadly present in mosquito genomes and that a large majority of the glycoproteins are physically associated with elements from the Pao family. We also found that most of the chuvirus-derived glycoproteins are structurally associated with potentially autonomous Pao elements and are likely to play a role in viral particle formation as an envelope protein. However, we also found structurally conserved solo glycoproteins, suggesting a potential role as an antiviral defense mechanisms.

MATERIALS AND METHODS

Data Collection

A non-redundant database of proteins including all taxa and a database of chuvirus genomes were obtained from NCBI (last accessed January 2018). Mosquito genomes were retrieved from NCBI and Vectorbase (last accessed January 2018) and chuvirus-derived EVEs already identified in mosquito genomes were retrieved from Whitfield et al. (2017). Mosquito genome sources and assembly metrics are presented in **Supplementary Material 1**. All command lines used in the present study can be found in **Supplementary Material 2**.

EVE Screening

Two BLAST-based strategies were used. The first used the chuvirus genome dataset as a query in a tBLASTx (Altschul et al., 1989) screening against mosquito genomes, and the second used the EVEs from Whitfield et al. (2017) as a query in a BLASTn analysis against mosquito genomes. All EVE regions were extracted from mosquito genomes in two ways: I—only the ungapped aligned region; II—ungapped aligned region plus 10 kb of each flanking region.

The resulting sequences were used as a query in a BLASTx analysis against the non-redundant protein database with different filters in order to eliminate false positives and false negatives chuvirus EVEs. Such filtering was performed in order to avoid two detected issues. First, as a number of viral proteins are still not annotated in the databases, if one considers the first hit alone in order to determine the viral origin one would disregard some EVEs. Second, some wrongly annotated viral proteins would cause the first hit to be miss-annotated as viral, while the following two or more hits would indicate that the sequence belongs to another taxon, generating a false positive result. Queries that showed four or fewer matches were annotated as an EVE if the best match was a viral protein. For queries with five or more subject matches, the proportion of the five best matches was taken into consideration, in accordance with the following criterion: a sequence is annotated as an EVE only if three or more subjects are viral proteins.

The EVE sequences containing the flanking sequences were clusterized using CD-HIT (Li and Godzik, 2005) to remove redundancy. Flanking sequences were retained to avoid clusterization of identical or very similar copies while allowing the removal of the same EVE copy recovered using the two search strategies (chuvirus genomes and chuvirus EVEs).

Chuvirus EVE Characterization

Using the above mentioned strategy, we obtained a total of 428 chuvirus-derived EVEs from mosquito genomes, of which 409 were glycoproteins, 18 RNA-dependent RNA polymerases and one nucleoprotein region. However, a number of these were found in small-size contigs and in close proximity to indeterminate "NNN" regions. In order to avoid genome assembly problems, we restricted our analysis to contigs bearing EVEs with at least 4 kb (Llorens et al., 2011) of each flanking region and no undetermined bases (**Supplementary Material 3**).

For the remaining 322 sequences, we extracted potential coding regions using the EMBOSS *getorf*[1] tool, and ORFs containing fewer than 100 amino acids were removed from further steps, resulting in 279 sequences.

Flanking Sequence Analysis

All EVEs plus flanking regions passing the aforementioned filters were translated using the EMBOSS *getorf* software, and the resulting ORFs were used in a domain-signature analysis with BATCH-CDD (Marchler-Bauer et al., 2011) to characterize the genomic context of each EVE locus. Sequences with BEL-Pao signature domains—GAG, PR, RT, RNAse H, and INT (whether or not flanked by long terminal repeats—LTRs)—were considered to be putative hybrids of a BEL-Pao element and chuvirus-derived sequences. For graphical representation, genetic maps are constructed with karyoploteR R-package (Gel and Serra, 2017). In order to identify orthologous EVE copies, we aligned EVEs plus flanking regions (around 20 kb of each side) with MAFFT (Katoh et al., 2001).

Search for Homologous BEL-Pao Elements

Nucleotide sequences of complete BEL-Pao elements containing domain signatures and LTRs recovered in the previous analysis were used as queries in a BLASTn analysis against the respective mosquito genomes to recover all homologous BEL-Pao copies. Sequences retrieved in this step were recovered with 10 kb of each flanking region and screened for the presence of chuvirus-derived EVEs. Long terminal repeat (LTR) regions from all

[1]http://www.bioinformatics.nl/cgi-bin/emboss/getorf

sequences retrieved were evaluated using LTR_FINDER with default parameters (Xu and Wang, 2007).

All EVEs were sorted, based on RdRp, Gly, or NP proteins and BEL-Pao Retrotransposons based on whether they are associated with chuvirus-derived glycoproteins or not. The structural conservation of the copies were sorted into the following categories: (i) potentially active retrotransposons containing both LTRs, BEL-Pao conserved protein domains (GAG-PR-RT-RH-INT) with or without a chuvirus-derived glycoprotein; (ii) defective elements which have at least one BEL-Pao domain and one or no LTRs; (iii) solo LTRs; and (iv) solo glycoproteins (**Table 1**).

Molecular Modeling

Molecular modeling was performed for three groups of glycoproteins: I—Solo glycoproteins, corresponding to glycoproteins without the BEL-Pao signature in their flanking regions; II—glycoproteins found inside of BEL-Pao element boundaries (LTRs)—in order to select some Gly proteins found inside BEL-Pao retrotransposons (group II), an amino acid distance matrix was created using UGENE with simple similarity distance algorithm maintaining gap regions (Okonechnikov et al., 2012). One random sequence was chosen for molecular modeling from each cluster exhibiting more than 90% similarity (**Supplementary Material 4**); and III—glycoproteins from *bona fide* chuviruses previously characterized by metatranscriptomics in mosquitoes.

The Phyre2 server (Kelley et al., 2015) was used to select a template for each protein of interest. The 3D structure of each template selected was obtained from the Protein Data Bank (PDB) (Berman et al., 2000) and submitted, together with the Gly sequence, to the Modeller package version 9.23 multimer modeling algorithm. The predicted models were evaluated for their stereochemical parameters using the Procheck tool (Laskowski et al., 1993).

Solo LTRs

LTRs from previously characterized elements were used as a query against mosquito genomes in a BLASTn analysis. Matches were recovered with 10 kb of the flanking regions. Sequences with a single LTR matching the mean length of LTRs identified using full BEL-Pao elements (around 670 base pairs) and no surrounding BEL-Pao domains were considered to be Solo LTRs (**Supplementary Material 5**).

Evolutionary Analysis

Two analyses were performed. The first was based on Gly protein sequences using only amino acid sequences with more than 100 amino acid residues. This analysis included most EVEs identified and chuvirus Gly from the literature and NCBI. The second analysis involved the evolutionary history of BEL-Pao retrotransposons and used amino acid sequences from reverse transcriptase and RNAse H (the most abundant domains present on BEL-Pao as characterized in the present study) (**Supplementary Material 6**). The second analysis used both BEL-Pao (with and without glycoprotein) and representative BEL-Pao retroelements from the five branches of the BEL-Pao superfamily included in the Gypsy Database 2.0[2].

The amino acid sequences for both analyses were aligned with the MAFFT algorithm (Katoh et al., 2001) and automatically edited using Gblocks (Castresana, 1999) with relaxed parameters (**Supplementary Material 2**). The most likely amino acid substitution model was determined using SMS (Lefort et al., 2017) on the ATCG online platform[3].

Phylogenetic trees reconstruction were carried out using MrBayes v3.2.2 × 64 (Ronquist et al., 2010), starting with three seeds and 1,000,000 generations. The convergence of the independent runs was detected when the standard deviation of all three seeds was lower than 0.05. The burn-in removed the first 25% of trees sampled and the remaining 75% were used to generate a posterior probability consensus tree and the phylogenies were visualized using iTOL (Letunic and Bork, 2007).

PCR Validation of Chuvirus Solo-Glycoproteins

Five solo glycoproteins without BEL-Pao signature were selected for PCR validation in three species available at the insectary of the Departamento de Entomologia—Instituto Aggeu Magalhães: *Aedes aegypti, Aedes albopictus,* and *Culex quinquefasciatus*, as well as the natural population of mosquitoes of the same species collected from different points at the Recife city (**Table 2**). The natural population of *Cx. quinquefasciatus* samples (Cxqui1301 and Cxqui1304) were collected in the Hospital das Clínicas—UFPE (8°02′51.9″S 34°56′45.6″W) during the years 2016 and 2017, as well as the sample Ae1471 of *Ae. aegypti*. The natural population samples of *Ae. albopictus*—AlbZoo and AlbJB—were collected in Parque Estadual Dois Irmãos (8°00′43.3″S 34°56′40.7″W) and Jardim Botânico do Recife (8°04′36.9″S 34°57′34.1″W), respectively, in 2017.

DNA extraction was performed from pools of 5 individuals (females and males) of each species with the protocol established by Ayres et al. (2003), the quality and concentration of DNA was evaluated with NanoDrop 2000 (Thermo Fisher Scientific). Primers to amplify the endogenous gene of the sodium channel (sodium channel protein para-like LOC109432678), were used as control to evaluate the DNA samples integrity before the EVE screening. The primers sets that amplify part of the EVE and a region of the mosquito genome were designed with Primmer3 (Koressaar and Remm, 2007) and validated with PrimmerBLAST (Ye et al., 2011) and OligoAnalyzer Tool (Owczarzy et al., 2008) against the mosquitoes reference genomes (**Figure 1A**). Primers information is shown in **Supplementary Material 7**. PCR was performed with GoTaq-Flexi G2 DNA-Polymerase following the Promega manufacturer's protocol. All PCR reactions are conducted in a final volume of 25 μL containing 1 μL of each primer (10 μM), 2uL of dNTP (0.2 mM each base), 5.0 μL of GoTaq Flexi Buffer, 4.0 μL of MgCl2 (25 mM), 0.25 μL of GoTaq-Flexi G2 DNA-Polymerase, 3.0 μL of DNA sample and 8.75 μL of Ultrapure H20. PCRs were conducted with the following

[2]http://gydb.org

[3]http://www.atgc-montpellier.fr/

TABLE 1 | General features of chuvirus-like proteins and BEL-Pao retroelements identified.

	Chuvirus endogenized regions				Anakin*				BEL-Pao retroelements without Glyco				
Genome	**Nucleop.**	**Solo-Glyco****	**Anakin***	**RdRp**	**Complete**	**Defective + LTR**	**Defective**	**Total**	**Complete**	**Defective + LTR**	**Defective**	**Total**	**Solo-LTR**
Aedes aegypti AaegL3	0	5/4	23	2	1	2	20	23	0	9	15	24	9
Aedes aegypti Aag2	1	15/8	55	4	0	20	35	55	0	9	9	18	14
Aedes aegypti BV	0	14/0	1	4	0	0	1	1	0	0	1	1	0
Aedes albopictus Rimini	0	20/0	1	2	0	0	1	1	0	0	0	0	0
Aedes albopictus Foshan	0	19/3	27	3	0	2	25	27	0	1	18	19	4
Aedes albopictus C636	0	3/3	85	3	9	15	61	85	2	20	28	50	82
Culex quinquefasciatus	0	3/3	17	0	3	9	5	17	0	0	0	0	20
Anopheles albimanus	0	0	3	0	0	0	3	3	0	0	0	0	0
Anopheles arabiensis	0	2/0	2	0	0	0	2	2	0	0	0	0	0
Anopheles atroparvus	0	4/0	1	0	0	0	1	1	0	0	0	0	0
Anopheles christyi	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles coluzzi	0	1/0	1	0	0	0	1	1	0	0	0	0	0
Anopheles culicifacies	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles darlingi	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles dirus	0	5/2	5	0	0	0	5	5	0	0	0	0	0
Anopheles epiroticus	0	4/0	3	0	0	0	3	3	0	0	0	0	0
Anopheles farauti	0	5/0	2	0	0	0	2	2	0	0	0	0	0
Anopheles funestus	0	7/1	1	0	0	0	1	1	0	0	0	0	0
Anopheles gambie PEST	0	0	2	0	0	0	2	2	0	0	0	0	0
Anopheles gambie Pim.	0	0	6	0	0	1	5	6	0	0	0	0	1
Anopheles gambie S	0	3/3	1	0	0	0	1	1	0	0	0	0	0
Anopheles gambie BV	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles koliensis	0	1/0	1	0	0	0	1	1	0	0	0	0	0
Anopheles maculatus m3	0	3/0	0	0	0	0	0	0	0	0	0	0	0
Anopheles maculatus BtQ1	0	12/3	15	0	0	1	14	15	0	0	0	0	4
Anopheles melas	0	1/0	0	0	0	0	0	0	0	0	0	0	0
Anopheles merus	0	1/1	1	0	0	0	1	1	0	0	1	1	0
Anopheles minimus	0	4/2	1	0	0	0	1	1	0	0	0	0	0
Anopheles nili	0	1/0	2	0	0	0	2	2	0	0	0	0	0
Anopheles punctulatus	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles quadriannulatus	0	3/0	2	0	0	0	2	2	0	0	0	0	0
Anopheles sinensis SIN.	0	1/1	1	0	0	0	1	1	0	0	0	0	0
Anopheles sinensis china	0	5/1	0	0	0	0	0	0	0	0	0	0	0
Anopheles stephensi Indian	0	1/1	1	0	0	0	1	1	0	0	2	2	0
Anopheles stephensi SDA	0	1/1	2	0	0	0	2	2	0	0	1	1	0
Anopheles cracens	0	0	0	0	0	0	0	0	0	0	0	0	0
Anopheles aquasalis	0	2/0	1	0	0	0	1	1	0	0	0	0	0
Total	1	146/37	263	18	13	50	200	263	2	39	75	116	134

**New retroviruses described in the present work, composed by Pao retrotransposons canonical structures plus Glycoprotein similar to Chuviridae Glycoproteins. **Non-filtred/filtred. Defective means elements that do not have all domains of a complete BEL-Pao retroelement (both LTRs, GAG protein and polyprotein with all domains).*

TABLE 2 | PCR validation of solo glycoprotein integration into the genome of different mosquito lineages.

	Ae. albopictus								*Ae. aegypti*				*Cx. quinquefasciatus*					
Sample	**AalIns***		**AalC636***		**AalZoo***		**AalJB***		**AaeIns***		**Aae1471***		**CxquiIns***		**Cxqui1304***		**Cxqui1331***	
Element/Flanking Region	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**	**L**	**R**
AalFosh_15**	+	−	+	−	+	−	+	−										
AalFosh_19**	+	−	−	−	+	−	+	−										
AAeliv_06_L**									−	+	−	+						
AegAag2_12**									+	+	+	+						
CquJoh_01**													−	+	−	+	−	+

*L/R columns represent left and right flanking regions screened by PCR, respectively. * Sample names, according to information present on section "PCR Validation of Chuvirus Solo-Glycoproteins". ** Element names, according to* ***Supplementary Material 3****.*

program: Initial denaturation at 94°C for 2 min, 45 cycles at 94°C for 1 min (denaturation), primer annealing at 51–59°C for 50 s (depending on specific TM of each primer pair), extension at 72°C for 1 min followed by a final extension at 72°C for 5 min. DNA amplification was visualized with 2% agarose gel stained with ethidium bromide and bands with expected were extracted, purified and sequenced with DNA ABI Prism 3100 Genetic Analyser (Applied Biosystems) from both forward and reverse strands.

Forward and reverse electropherograms are analyzed with Geneious Prime version 2019.1.3 (Kearse et al., 2012) and consensus sequences were generated. Contigs from sequenced products are then aligned with EVEs identified by *in silico* analyses with Aliview (Larsson, 2014).

RESULTS

Supporting Evidence of Widespread Chuvirus Endogenization in Different Mosquito Genomes

Four hundred and twenty-eight EVEs were identified in 32 out of 37 genomes screened. These elements correspond to 409 glycoprotein fragments, 18 RNA-dependent RNA-polymerases (RdRp) and one nucleoprotein (**Supplementary Material 8**). After the exclusion of EVEs from small contigs and those in close proximity to uncertain sequences/assembled bases (NNNs), 279 sequences remained. Two hundred forty-one of these presented glycoprotein conserved amino acid domains.

In view of the previously described and confirmed abundance of chuvirus-derived glycoprotein EVEs, a more in-depth characterization of these EVEs was performed. Endogenous glycoproteins varied in size from 117 to 1977 nucleotides (median = 880) and amino acid length from 39 to 659 residues. The amino acid identity with each chuvirus genome used as a query varied from 29.03 to 56.41% (average = 32.83, *SD* = 6.31), the alignment length varied from 35 to 910 amino acids (average = 283.77, *SD* = 191.16) and *e*-value varied from 9,00E-05 to 0.0 (average = 6,00E-07, *SD* = 6,00E-06) (**Supplementary Material 8**). We could not identify orthologous EVE copies based on flanking region alignment analyses.

Glycoproteins Are Mostly Associated With Elements From the BEL-Pao Retrotransposon Superfamily

Screening for BEL-Pao protein domain signatures and LTRs we found glycoproteins in three different contexts (**Figures 1**, **2** and **Table 1**). Thirty-seven glycoproteins were categorized as solo glycoproteins, since no BEL-Pao protein domain or LTR signature was found (**Figure 1A**). Two hundred glycoproteins were flanked by BEL-Pao protein domains but with only one LTR or none, characterizing them as defective BEL-Pao elements (**Figure 1B**). Sixty-three glycoproteins were associated with BEL-Pao domains flanked by two LTRs, characterizing them as potentially active elements. Thirteen of the BEL-Pao elements including chuvirus-like glycoproteins had all domains and complete LTRs. We call these elements "Anakin." This name refers to the putative "switching sides" of viral glycoprotein found into a potential new retrovirus and solo glycoprotein that is likely counteracting against this retrovirus and circulating Chuviruses (Anakin element, **Figure 1C**).

We also identified solo LTRs in seven genomes (**Table 1**). These vary in number from only one in the *Anopheles gambiae* Pimpirena assembly to 82 solo LTRs in the *Aedes albopictus* C636 assembly (**Supplementary Material 5**). Interestingly, the number of solo LTRs was found to be greater than the number of LTRs associated with full BEL-Pao elements in the *Aedes albopictus* C636 and *Anopheles maculatus* BTQ1 assemblies (**Table 1**).

Endogenous and Exogenous Glycoproteins Have Similar 3D Structures

Although BLAST and phylogenetic analysis indicate that the EVEs found share a common ancestor with chuviruses, the amino acid identity is considerably low (between 25 and 50%). Another way to obtain further support for the viral origin of these endogenous sequences is through 3D structure modeling. If these polypeptides resemble viral glyco/envelope proteins in 3D space or have similar protein domains (Malik and Henikoff, 2005), this would corroborate their viral origin, as well as the protein function (Webb and Sali, 2017).

The templates identified by the Phyre2 tool represent many type B glycoproteins homotrimers from different

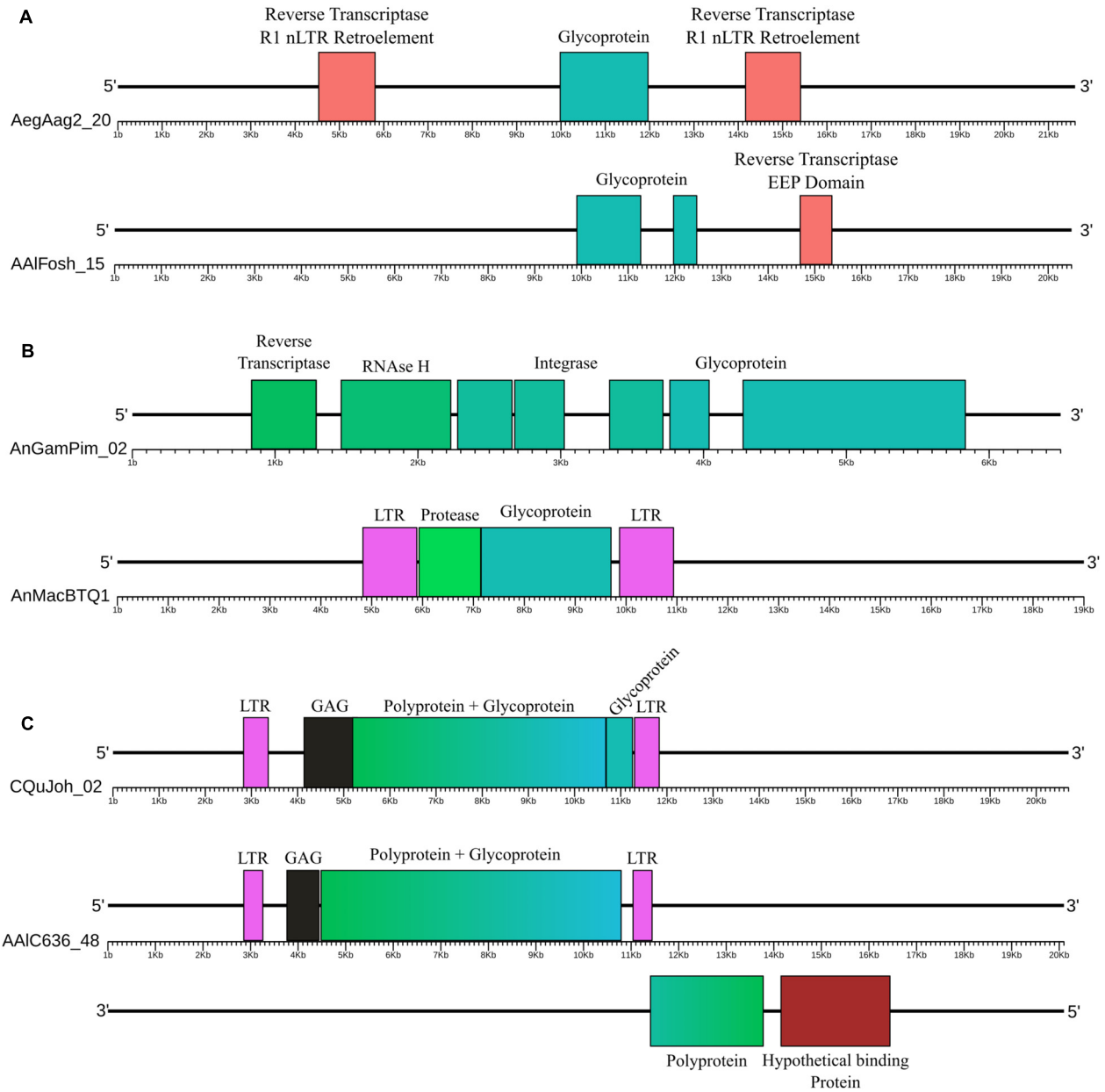

FIGURE 1 | Examples of glycoproteins from the *Chuviridae* family. **(A)** EVEs not associated with BEL-Pao Retroelements, AegAag2_20 of the *Ae. aegypti* genome (Aag2 assembly) and AAlFosh_15 of the *Ae. albopictus* genome (AaloF1 assembly); **(B)** EVEs associated with non-autonomous BEL-Pao retroelements, AnGamPim_02 of the *An. gambiae* genome (assembly Pimpirena S) and AnMacBTQ1 of the *An. maculatus* genome (BtQ1 assembly); and **(C)** EVEs associated with complete BEL-Pao retroelements (Anakin), CQuJon_02 of the *Cx. quinquefasciatus* genome (CpipJ2 assembly) and AALC636_48 of the *Ae. albopictus* genome (Canu80x assembly). Reddish rectangles are protein domains from other non-BEL-Pao retrotransposons; Blue and green reactangles are glyproteins, integrases, RNAse H and RT domains; Black rectangles are capsid domains; Pink rectanges are retrotransposon LTRs and Dark red rectangles are other hypothetical domains.

viruses (**Figure 3**). It was possible to reconstruct three-dimensional models for all glycoproteins analyzed. All pdb files of modeled glycoproteins are available on **Supplementary Material 9** and Ramachandran Plots region values can be seen in **Supplementary Material 10**. Only two 3D models (AnMacBtQ1_01 and Wuhan Mosquito Virus 8) had more than 1.0% residues in disallowed regions. For all glycoproteins modeled the residues in core regions varied between 82.3 and 88.3% and in allowed regions from 9.3 to 14.2%.

The TM-alignment between representatives of solo glycoproteins, glycoproteins fused with BEL-Pao elements and *bona fide* chuvirus glycoproteins demonstrates a similarity of three-dimensional structures greater than 0.7 (**Figure 3B**),

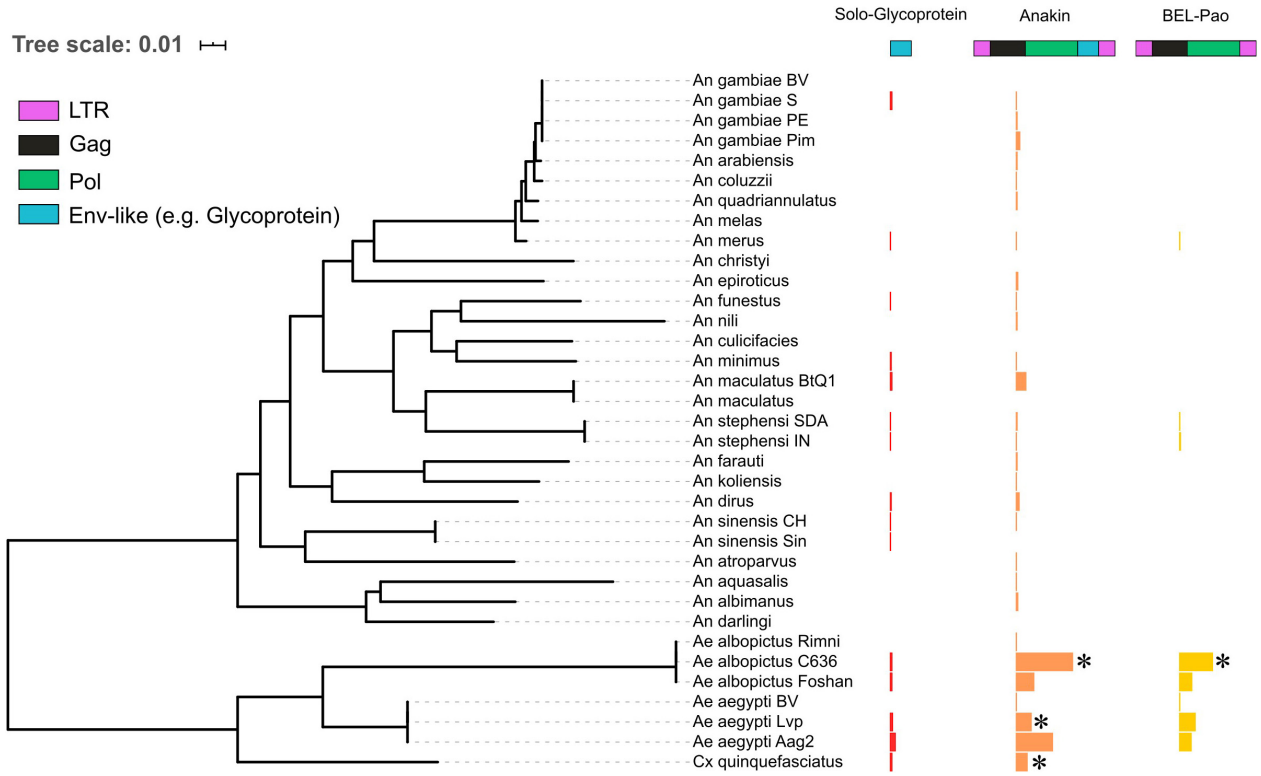

FIGURE 2 | The distribution of solo endogenized glycoproteins, Anakin retroelements and BEL-Pao retrotransposons without glycoprotein across mosquito genomes. Horizontal bars width denotes the copy number of each elements characterized: Red, Solo-Glycoprotein; Orange, Anakin; and Yellow, BEL-Pao retrotransposon without chuvirus glycoprotein. *Denotes genomes with complete elements having both LTR and all retrotransposons domains.

providing strong evidence that these glycoproteins are folded in a similar way.

Evidence of a Chuvirus ENV-Like Protein Captured by Elements From the BEL-Pao Superfamily

Bona fide chuviruses with non-segmented genomes (sequences in yellow) form a basal clade in the glycoprotein evolutionary tree (**Figure 4**), confirming the common origin of EVE glycoproteins and non-segmented chuviruses. It is also worth noting the existence of a basal clade of *Aedes aegypti* EVEs clustered with one *bona fide* chuvirus—Mos8Chu0 (**Figure 5A**).

Four main findings can be reported regarding EVE clades: (i) the presence of various high-identical EVE copies, indicated by clades containing several sequences with near-zero branch lengths (**Figures 5B,D**); (ii) the presence of glycoproteins inside BEL-Pao retroelements from RepBase intervening into various EVE clades (**Figure 5C**); (iii) the presence of several solo-glycoproteins embedded into retroelements clades (**Figure 5B**), and (iv) the presence of sequences described as *bona fide* chuviruses (Lara Pinto et al., 2017; Pauvolid-Corrêa et al., 2016) inside and outside of clades mostly composed of EVE copies (**Figure 5D**).

The evolutionary tree using the RT and RH regions from both BEL-Pao retroelements, including elements containing or not chuvirus glycoproteins, shows four clearly defined clades (**Figure 6**). One of these has 181 elements without glycoproteins closely related to BEL elements in Branch 1 (**Figure 6**). The other three are closely related clades with Pao elements from Branch 2, of which 129 are composed of elements with glycoproteins (red scalene triangles) and 122 of elements without glycoproteins (blue scalene triangles, **Figure 6**). It is clear that chuviruses glycoproteins are specifically associated with elements from the Pao family—the now called Anakin elements.

There is a large number of Pao elements without glycoproteins, but comparing it with the Anakin elements the last are more abundant and widespread, in a larger number of mosquito species, than Pao elements without glycoproteins (**Figure 2** and **Supplementary Material 11**).

PCR Validation of Solo-Glycoproteins in Different Mosquito Strains

Five primer pairs were designed to amplify both chuvirus/mosquito integration junction of solo-glycoprotein regions. Four lineages of *Ae. albopictus*, two lineages of *C. quinquefasciatus* and two lineages of *Ae. aegypti* were screened (**Table 2** and **Supplementary Material 12**). At least one EVE-mosquito boundaries were sequenced successfully for all five solo glycoproteins (**Table 2**), showing a consistent alignment to the corresponding genomic region in the available genomes

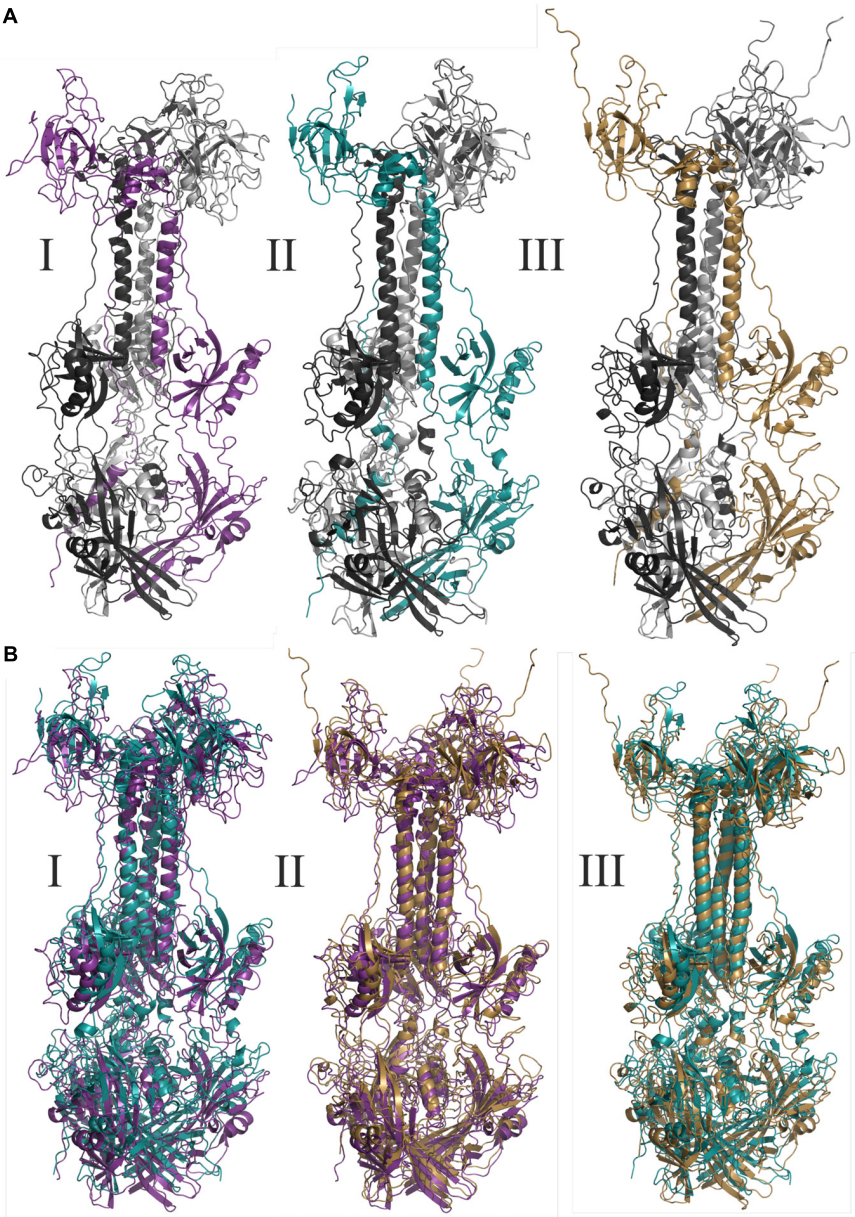

FIGURE 3 | Molecular modeling of glycoproteins. **(A)** Tridimensional models where **I** AegAag2_20 element, an example of solo-glycoprotein; **II** AAIC636_23 element, an example of glycoprotein fusioned with a complete Pao element; and **III** represents the glycoprotein of Mos8Chu0. **(B)** Represents the tridimensional glycoproteins B alignments between: **I:** AegAag2_20 in purple and AAIC636_23 in blue, with TM-score equals to 0.79415; **II:** AegAag2_20 in purple and Mos8Chu0 in yellow, with TM-score equals to 0.76639; and **III:** AAIC636_23 in blue and Mos8Chu0 in yellow, with TM-score equals to 0.80812.

(**Supplementary Material 12**) confirming the integration of the EVEs and highlighting that these insertions were conserved and an ancient component of these genomes.

DISCUSSION

Viruses and transposable elements share a common evolutionary history and hence shared several features, particularly for some retroviruses and retrotransposons that are differentiated by the presence/absence of an envelope gene (glycoprotein). The envelope protein is responsible for the infection capacity of the former (Hayward, 2017). But more than that, these entities exchange genetic information between them and with their host genome (Drezen et al., 2017; Frank and Feschotte, 2017; Gilbert and Cordaux, 2017; Sinha and Johnson, 2017). Several instances of exchange of functional genes that allowed major changes in their evolutionary history are known such as the acquisition of a complete herpesvirus by a piggyBac transposon (Inoue et al., 2017) or the emergence of retroviruses by envelope capture

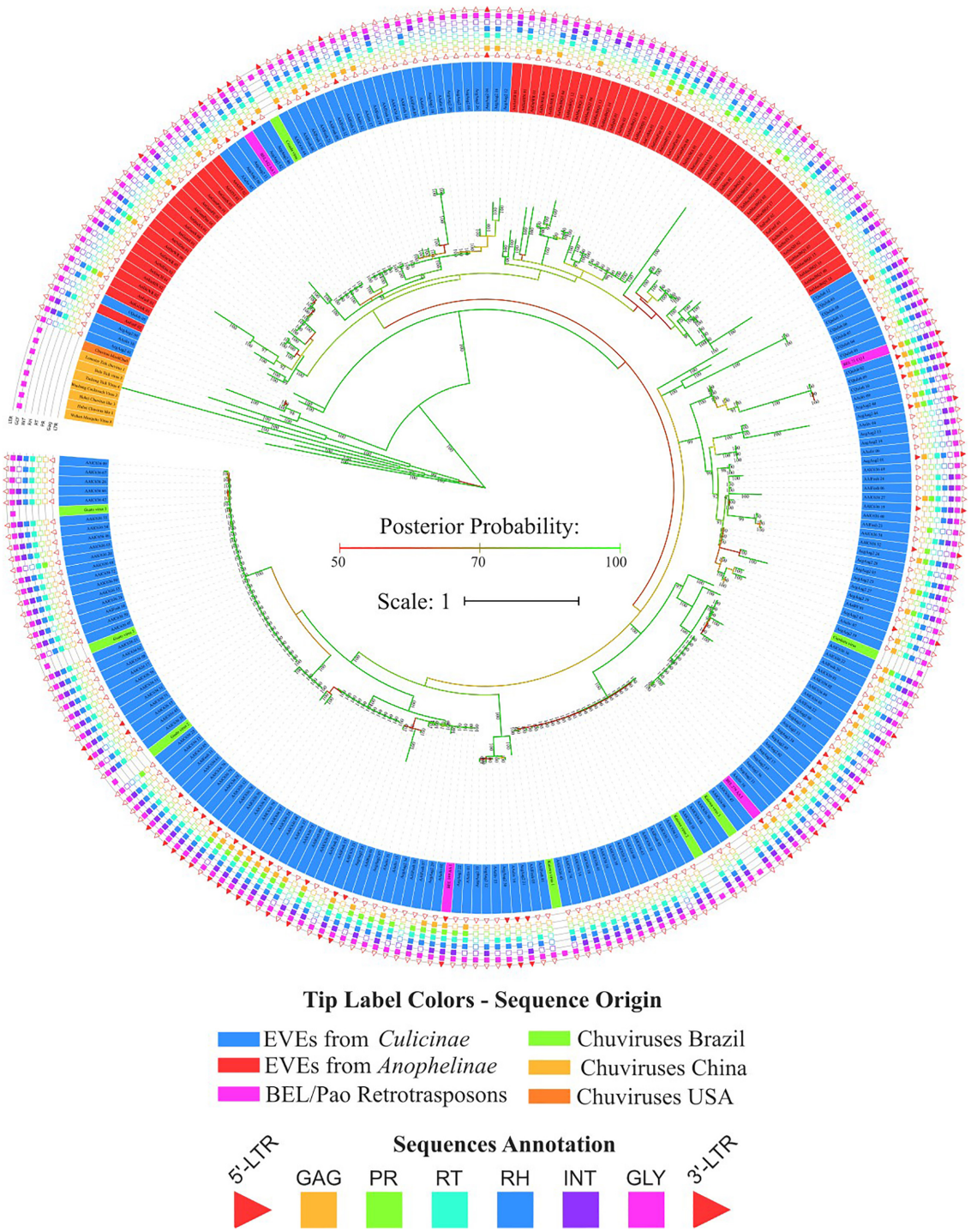

FIGURE 4 | Phylogenetic reconstruction of glycoproteins from *Chuviridae* family. Bayesian tree reconstructed after 5,000,000 generations from 3 seeds with standard deviation mean between final trees equal to 0.02. Posterior probability are depicted over each node, only values greater than 90 were plotted; Tip label colors denotes: Sequences identified as EVEs derived from *Chuviridae* family, blue and red color represents EVEs identified in *Culicinae* and *Anophelinae* subfamilies, respectively; Pink labels are retrotransposons of BEL-Pao superfamily available on RepBase and that showed similarity with chuviruses glycoproteins; Light orange labels are chuviruses identified from samples of different arthropods sampled in China by Shi et al. (2016); green labels are sequences described in Brazil as chuviruses by Lara Pinto et al. (2017); dark orange are chuvirus available on NCBI (access KX924631.1). Sequence structure: LTR, Long Terminal Repeat—red triangle; GAG Capsid protein—orange square; PR, Protease—light green square; RT, Reverse Transcriptase—light blue square; RH, RNAse H—blue square; Integras—purple square and GLY, Glycoprotein—pink square. Phylogeny available on: https://itol.embl.de/tree/20013326132982156353869 3.

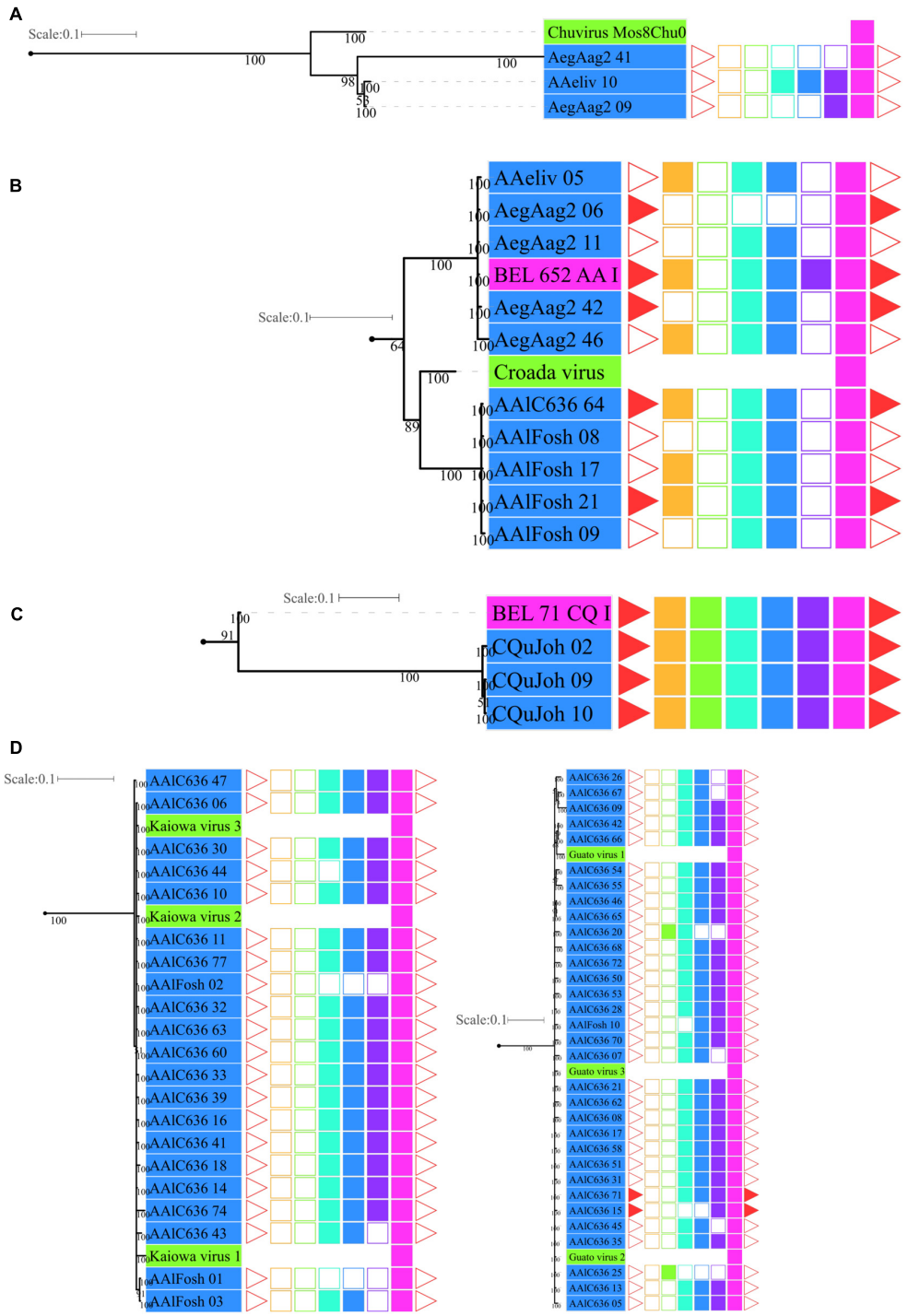

FIGURE 5 | Phylogenetic clades showing particular clustering of some sequences from **Figure 4**. **(A)** Basal clade with *bona fide* chuvirus, chuvirus solo-glycoprotein and BEL-Pao derived elements associated with chuvirus glycoprotein; **(B)** clades with a number of highly similar glycoproteins found associated with BEL-Pao elements; **(C)** clade with complete BEL-Pao Retrotroelements (Anakin); **(D)** clades with BEL-Pao derived elements associated with chuvirus glycoprotein and potential chuvirus sequences described by Lara Pinto et al. (2017). Sequences identified as EVEs derived from *Chuviridae* family, blue and red color range labels represent EVEs identified in *Culicinae* and *Anophelinae* subfamilies, respectively, retrotransposons of BEL-Pao superfamily available on RepBase and that showed similarity with chuviruses glycoproteins, range labels colored with purple; Chuviruses identified in China by Shi et al. (2016), range labels colored with light orange; Sequences described in Brazil as chuviruses by Lara Pinto et al. (2017), range labels colored with light green; Chuvirus available on NCBI (access KX924631.1), range label colored with dark orange; LTR, Long Terminal Repeat—red triangle; GAG, Capsid protein—orange square; PR, Protease—light green square; RT, Reverse Transcriptase—light blue square; RH, RNAse H—blue square; Integrase—purple square and GLY, Glycoprotein—pink square.

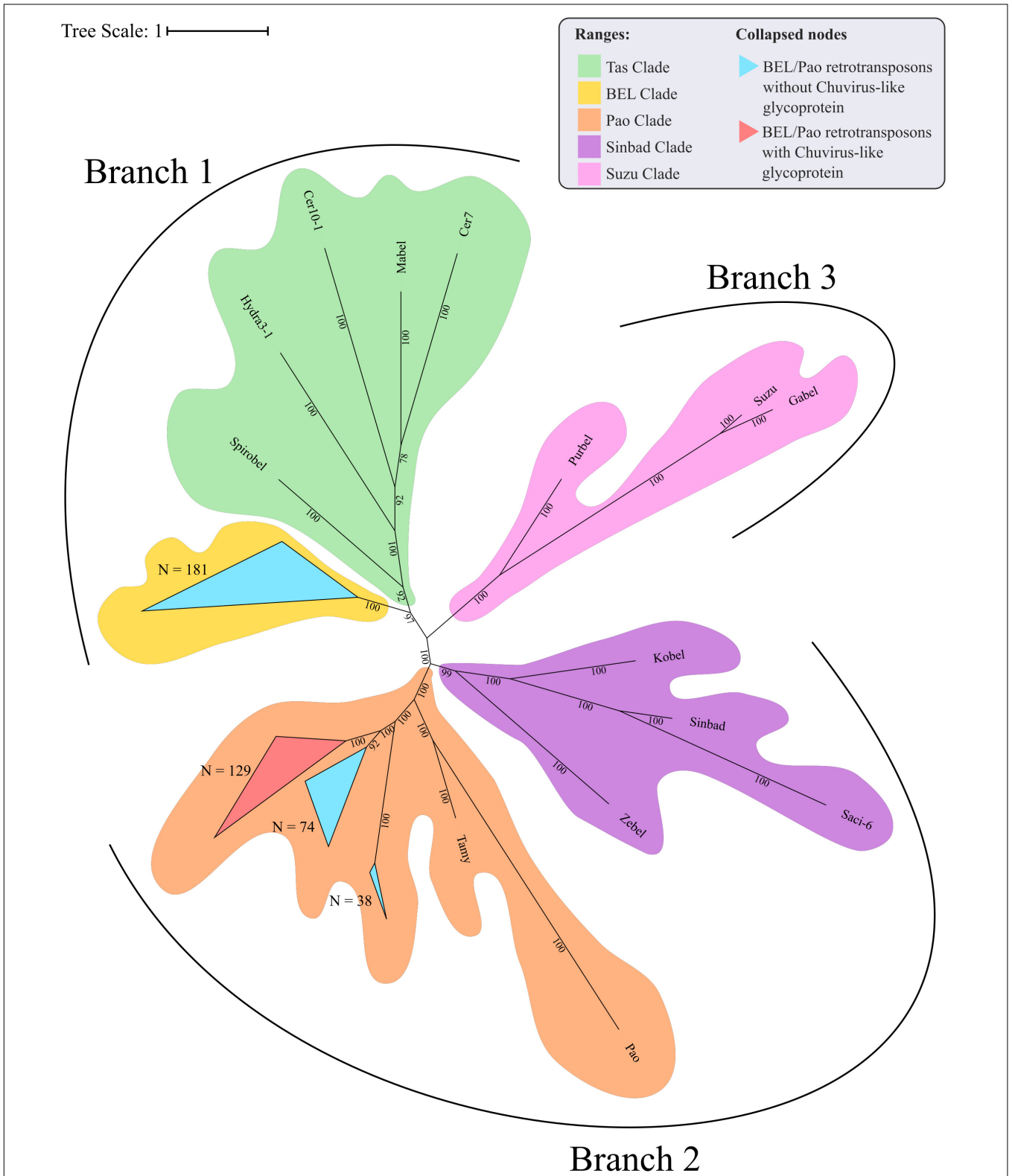

FIGURE 6 | Phylogenetic representation of Reverse Transcriptase and RNAse H regions of Polyproteins from BEL-Pao Retroelements. Bayesian tree constructed after 1,000,000 generations from 3 seeds with standard deviation mean between final trees equal to 0.03. Support values in posterior probability. Branches and clades are annotated according to BEL/Pao Retroelements phylogeny available on: http://gydb.org/index.php/. Uncollapsed phylogeny with domain annotation is present on https://itol.embl.de/tree/20013326140558156779232.

(Kim et al., 2004). Here, we characterized a new event of gene sequence exchange between viruses, the mosquito host genome and retrotransposons with the acquisition of a glycoprotein (envelope) gene from a chuvirus by a Pao retrotransposon (named Anakin) suggesting a past and likely current arms race between these entities.

The *Chuviridae* family was first described by Shi et al. (2016) in several arthropod species (including mosquitoes) through metatranscriptomic sequencing. Chuviruses have three possible genomic structures: a complete or bi-segmented circular genome, identified in ticks, crab, flies, spiders, cockroaches and mosquitoes, and a linear structure, identified in flies and crabs. In the bi-segmented structure, the glycoprotein gene is always flanked downstream by a nucleoprotein and by a viral particle protein (Shi et al., 2016). In the same study, the authors identified endogenous chuvirus elements in both mosquito and other insect genomes. Two subsequent studies also identified EVEs derived from chuvirus in mosquito genomes (Whitfield et al., 2017; Russo et al., 2019) but focused only on the *Aedes aegypti* genome. Whitfield et al. (2017) characterized two interesting features of chuvirus-derived EVEs from *Ae. aegypti*: (i) large differences among endogenized genomic regions, with several glycoproteins and few RdRps and nucleoproteins; and (ii) enrichment of BEL-Pao retroelements around these EVEs (Whitfield et al., 2017; Russo et al., 2019).

Virus genome replication is tightly linked to the virus genome structure. Replication origin and orientation may favor the endogenization of the genomic regions that are first copied/transcribed, since these regions are produced more abundantly than the last replicated/transcribed regions (Whitfield et al., 2017; Russo et al., 2019). Although little is known about chuvirus replication, the position of the genes in non-segmented genomes suggests that endogenization should occur more frequently for nucleoproteins or RdRp (in the terminal regions of the virus genomes) than for glycoproteins (around the middle of the virus genome). However, for segmented genomes, each segment could be integrated independently and hence one should expect to find a similar amount of different integrated viral genomic regions. We detected more glycoproteins than nucleoproteins and RdRps regions integrated into the mosquito genomes. This endogenization pattern is not expected to any segmented or non-segmented Chuvirus genomes and hence does not allow us to reach a conclusion about the genomic structure of the original chuvirus genome. But, some additional evidence points to another more likely explanation to explain the glycoprotein discrepancy. We detected that many of these glycoproteins are integrated into potentially active BEL-Pao retrotransposons, more specifically into retroelements from the Pao family (**Figure 6**). It is important to note that a recent preprint partially confirmed these findings of chuvirus EVEs into the LTR boundaries of BEL-Pao elements (Crava et al., 2020). In addition, the phylogenetic analysis showed that both glycoprotein associated or not with Pao retrotransposons domains are embedded into multiple phylogenetic supported clades of Pao retroelements (**Figure 4**) and a similar amount of solo glycoproteins and other solo Pao derived domains (19 RH-RT) were found scattered into the mosquito genomes suggesting that the high glycoprotein copy number is a result of the replication of Pao retrotransposons with posterior decay of Pao protein domains and that this protein may be an integral part of these elements as an envelope gene. It is important to note that most of the findings reported in this study were derived from genome sequences of cultured cells of *Ae. albopictus* (C6/36) and *Ae. aegypti* (Aag2) and lab-reared strain from *Ae. aegypti* (Lvp) that have been maintained in the lab for many years, therefore further investigation on wild mosquito populations are necessary to access the evolutionary implications of the findings in natural populations.

Acquisition and loss of envelope proteins have been detected in a number of viruses and retrotransposons, blurring the distinction between these entities. Substantial evidence already exists showing that many retroviruses can turn into an intragenomic lifestyle when the ENV protein is lost or becomes defective as a result of mutations and that various retroviruses have originated from retrotransposons (intragenomic lifestyle) that acquired new envelope genes (Malik et al., 2000). Specifically for BEL-Pao retrotransposons, there are three well-studied examples of ENV-like protein acquisition by active retroelements (Malik et al., 2000). The *Tas* retrotransposon from *Ascaris lumbricoides* acquired an ENV protein from *Phlebovirus* (Felder et al., 1994). The *Cer* retrotransposon from *Caenorhabditis elegans* acquired an ENV protein from Herpesvirus (Browning et al., 1996) and an ENV was acquired from a Gypsy retrovirus by a Roo retrotransposon (Malik and Henikoff, 2005). We describe here a fourth event, involving Pao retrotransposons and glycoproteins of chuvirus, supported by the strong phylogenetic association between glycoproteins of Pao retrotransposons and EVEs from the *Chuviridae* family and the identification of glycoproteins inside complete Pao structures flanked by LTRs. This recombination event associated with further retrotransposition of Anakin explains the high abundance of glycoproteins inside mosquito genomes when compared with the other chuvirus proteins.

The distribution of EVEs derived from chuvirus in mosquito species from both the Culicinae and Anophelinae subfamilies dates the integration of chuvirus glycoproteins into the ancestor of the Culicidae family, around 190 million years ago (Hedges et al., 2006). The endogenization of a chuvirus glycoprotein may have occurred directly into a Pao retrotransposon, thereby giving rise to the Pao retrovirus (**Figure 7A**). Alternatively, this hybrid element may have emerged from a recombination event involving a Pao retroelement and a chuvirus-derived glycoprotein after its endogenization (**Figure 7A**). After viral envelope protein endogenization into the host genome, two major events may occur: exaptation or molecular domestication of the viral protein leading to a new host function or the emergence of an antiviral mechanism. The domesticated viral envelope may be selected as a countermeasure against the cognate circulating virus to effectively prevent new virus particles from entering the cell. For instance, host endogenous envelope proteins may be produced and exported to the extracellular space by the host cells. These polypeptides will then compete for the host cell receptors with circulating viruses (Robinson et al., 1981; Ito et al., 2013; Armezzani et al., 2014;

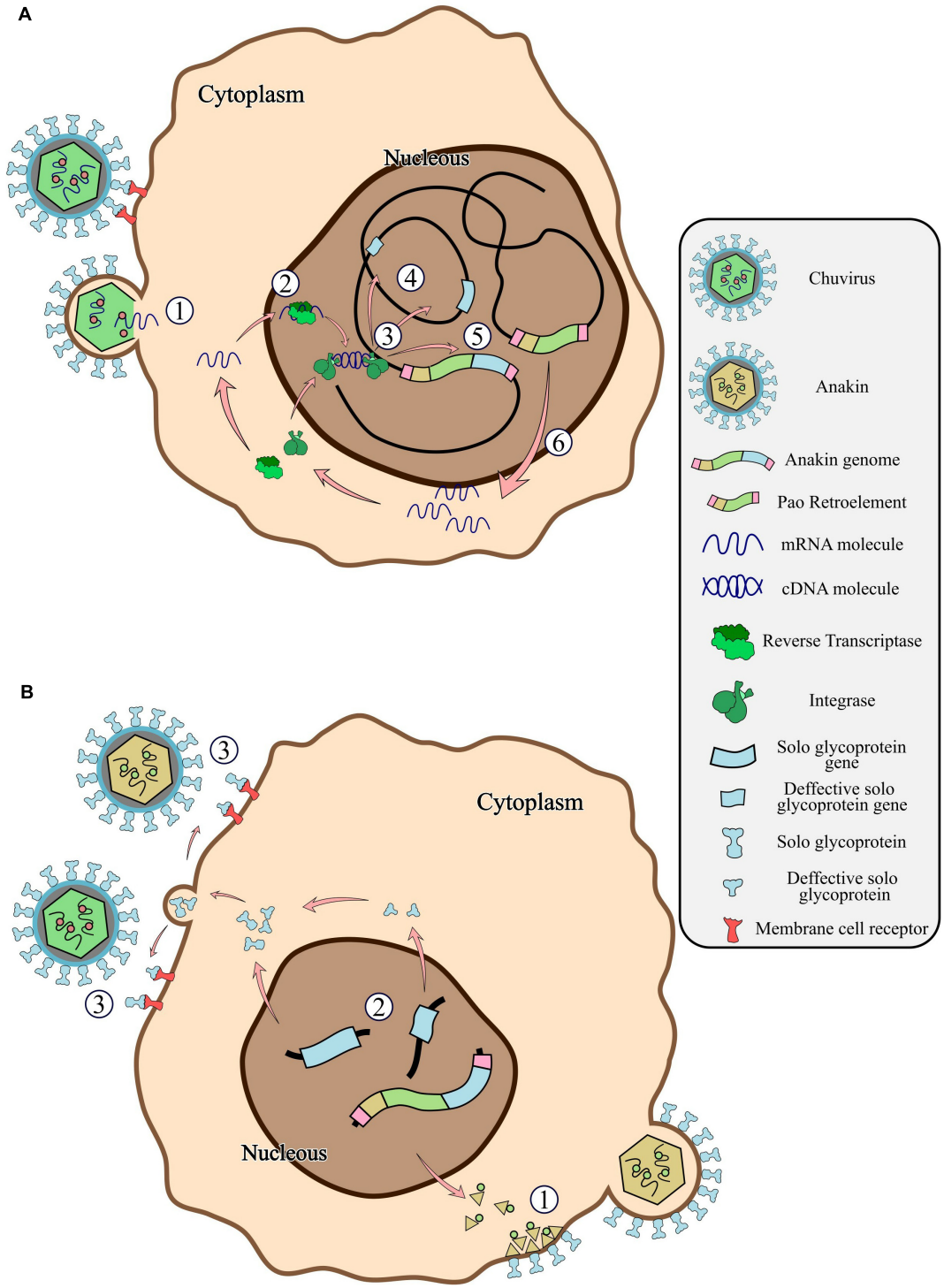

FIGURE 7 | General scheme of the chuvirus derived EVE integration into the germinative cells of the host genome, fusion with *Pao* retroelements and potential antiviral mechanisms by receptor competition. **(A)** Bonafide circulating chuvirus infecting a host cell and releasing their single-strand RNA into the cell cytoplasm (1); RT from retrotransposons fortuitously recognize the viral genome and retro transcribe to dsDNA (2); dsDNA is directly integrated into the host genome by double-strand break repair mechanisms or by recognition and integration using integrase proteins from retroelements (3) either in different genomic loci (4) or into a *Pao* retroelement boundaries (5); Expression of Pao proteins (6); **(B)** chuvirus-derived glycoproteins are dispersed in many genomic loci but have been replicated substantially by the *Pao* + chuvirus glycoprotein element (Anakin). Anakin retrovirus can assemble viral particles with envelope proteins allowing it to infect new host cells (1); Complete or defective solo glycoprotein are translated and exported to the extracellular environment (2) and can block new viral infection of both Anakin and chuvirus by binding to cell membrane virus protein receptors (3).

Johnson, 2019). The similarity of solo glycoprotein and glycoproteins from *bona fide* chuviruses and the Anakin retrovirus, as shown by sequence identity (**Supplementary Material 13**) and comparison of three-dimensional structures (**Figure 5A**), indicates that some of these endogenous glycoproteins may play a role in antiviral defense mechanisms against circulating chuviruses or even against Anakin retroviruses through competition for cell receptors (**Figure 7B**). This hypothesis is corroborated by the conservation of several solo glycoprotein integrations in different populations of *Ae. aegypti*, *Ae. Albopictus,* and *Cx. quinquefasciatus* species (**Table 2**). But, functional experiments are warranted to evaluate the transcription/translation of EVEs, its role as a receptor-blocker and the production of chuvirus and Anakin virus particles.

Finally, our results show the importance of taking EVEs into account in metaviromic studies (Nouri et al., 2018). Some of the chuvirus EVEs detected in our study are highly identical at the nucleotide level (99.08–100%) with previously described circulating chuviruses *Kaiowa, Guato, Cumbaru,* and *Croada* (Lara Pinto et al., 2017; Pauvolid-Corrêa et al., 2016). Although, we cannot rule out the possibility that some of these sequences may indeed have come from circulating chuviruses that infect *Ae. albopictus* species, the high identity values of these "viruses" with the EVEs found here strongly suggest that these previously defined chuviruses are in fact transcribing endogenous elements from the *Ae. albopictus* genome (**Supplementary Material 14**).

In this study we revealed the diversity of endogenous virus elements derived from the *Chuviridae* family in mosquito genomes. Our results show that such EVEs are widely distributed across the *Culicidae* family and are possibly involved in two major processes: the replication of Pao retroelements, through the acquisition of chuvirus glycoproteins; and a possible antiviral response against *bona fide* chuviruses and Anakin retroviruses originating from a fusion of Pao retroelements and the chuvirus envelope gene. These results shed new light on the dynamic evolution of EVEs and retrotransposons inside the mosquito genomes and point to the need for *in vitro* and *in vivo* experiments to test the hypothesis raised above using different mosquito populations, considering that our results are mainly based on cell culture assembly genomes.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Figshare, https://figshare.com/articles/In_and_outs_of_Chuviridae_endogenous_viral_elements_origin_of_a_retrovirus_and_signature_of_ancient_and_ongoing_arms_race_in_mosquito_genomes/11336258.

AUTHOR CONTRIBUTIONS

GW conceived the study. GW and AR planned and supervised the work. FD carried out the genomics analysis and wrote the manuscript. CV carried out the molecular modeling analysis. All authors agreed with the final version of the manuscript.

FUNDING

This work was supported by the Fundação de Amparo à Pesquisa do Estado de Pernambuco (FACEPE), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under the project number 406667/2016-0 and for the research grant PQ-2 of Wallau, GL (303902/2019-1).

ACKNOWLEDGMENTS

We thank the bioinformatics core at the Instituto Aggeu Magalhães (IAM) for technical assistance. We thank the collaborators from the Departamento de Entomologia—Instituto Aggeu Magalhães for the wild mosquitoes samples. This manuscript has been released as a pre-print at biorxiv (Dezordi et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.542437/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1989). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Armezzani, A., Varela, M., Spencer, T. E., Palmarini, M., and Arnaud, F. (2014). "Ménage à Trois": the evolutionary interplay between JSRV, enJSRVs and domestic sheep. *Viruses* 6, 4926–4945. doi: 10.3390/v6124926

Ayres, C. F. J., Melo-Santos, M. A. V., Solé-Cava, A. M., and Furtado, A. F. (2003). Genetic differentiation of *Aedes aegypti* (Diptera: Culicidae), the major dengue vector in Brazil. *J. Med. Entomol.* 40, 430–435. doi: 10.1603/0022-2585-40.4.430

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bha, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acid Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Browning, H., Berkowitz, L., Madej, C., Paulsen, J. E., Zolan, M. E., and Strome, S. (1996). Macrorestriction analysis of *Caenorhabditis elegans* genomic DNA. *Genetics* 144, 609–619.

Castresana, J. (1999). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Crava, C., Varghese, F. S., Pischedda, E., Halbach, R., Palatini, U., Marconcini, M., et al. (2020). Immunity to infections in arboviral vectors by integrated viral sequences: an evolutionary perspective. *Biorxiv* [Preprint]. doi: 10.1101/2020.04.02.022509v1

Dezordi, F. Z., Vasconcelos, C. R. S., Rezende, A. M., and Wallau, G. L. (2020). In and outs of Chuviridae endogenous viral elements: origin of a retrovirus and signature of ancient and ongoing arms race in mosquito genomes. *Biorxiv* [Preprint]. doi: 10.1101/2020.02.15.950899v1

Drezen, J.-M., Leobold, M., Bézier, A., Huguet, E., Volkoff, A.-N., and Herniou, E. A. (2017). Endogenous viruses of parasitic wasps: variations on a common theme. *Curr. Opin. Virol.* 25, 41–48. doi: 10.1016/j.coviro.2017.07.002

Felder, H., Herzceg, A., de Chastonay, Y., Aeby, P., Tobler, H., and Müller, F. (1994). Tas, a retrotransposon from the parasitic nematode *Ascaris lumbricoides*. *Gene* 149, 219–225. doi: 10.1016/0378-1119(94)90153-8

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296. doi: 10.1038/nrg3199

Forterre, P., and Prangishvili, D. (2009). The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. N.Y. Acad. Sci.* 1178, 65–77. doi: 10.1111/j.1749-6632.2009.04993.x

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

Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 33, 3088–3090. doi: 10.1093/bioinformatics/btx346

Gilbert, C., and Cordaux, R. (2017). Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr. Opin. Virol.* 25, 16–22. doi: 10.1016/j.coviro.2017.06.005

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., and Boulesteix, M. (2015). De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biol. Evol.* 7, 1192–1205. doi: 10.1093/gbe/evv050

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Curr. Opin. Virol.* 25, 23–27. doi: 10.1016/j.coviro.2017.06.006

Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971–2972. doi: 10.1093/bioinformatics/btl505

Inoue, Y., Saga, T., Aikawa, T., Kumagai, M., Shimada, A., Kawaguchi, Y., et al. (2017). Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish. *Nat. Commun.* 8:551. doi: 10.1038/s41467-017-00527-2

Ito, J., Watanabe, S., Hiratsuka, T., Kuse, K., Odahara, Y., Ochi, H., et al. (2013). Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. *J. Virol.* 87, 12029–12040. doi: 10.1128/JVI.01267-13

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* 17, 355–370. doi: 10.1038/s41579-019-0189-2

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2001). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kim, F. J., Battini, J.-L., Manel, N., and Sitbon, M. (2004). Emergence of vertebrate retroviruses and envelope capture. *Virology* 318, 183–191. doi: 10.1016/j.virol.2003.09.026

Koressaar, T., and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23, 1289–1291. doi: 10.1093/bioinformatics/btm091

Lara Pinto, A. Z., de Santos de Carvalho, M., de Melo, F. L., Ribeiro, A. L. M., Morais Ribeiro, B., and Dezengrini Slhessarenko, R. (2017). Novel viruses in salivary glands of mosquitoes from sylvatic Cerrado, Midwestern Brazil. *PLoS One* 12:e0187429. doi: 10.1371/journal.pone.0187429

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). *PROCHECK*: a program to check the stereochemical quality of protein structures. *J. Appl. Crystal* 26, 283–291. doi: 10.1107/S0021889892009944

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128. doi: 10.1093/bioinformatics/btl529

Li, W., and Godzik, A. (2005). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70–D74. doi: 10.1093/nar/gkq1061

Malik, H. S., and Henikoff, S. (2005). Positive selection of Iris, a retroviral envelope-derived host gene in *Drosophila melanogaster*. *PLoS Genet.* 1:e0010044. doi: 10.1371/journal.pgen.0010044

Malik, H. S., Henikoff, S., and Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* 10, 1307–1318. doi: 10.1101/gr.145000

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Nene, V., Wortman, J., Lawson, D., and Haas, B. (2006). Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316, 1718–1723. doi: 10.1126/science.1138878

Nouri, S., Matsumura, E. E., Kuo, Y.-W., and Falk, B. W. (2018). Insect-specific viruses: from discovery to potential translational applications. *Curr. Opin. Virol.* 33, 33–41. doi: 10.1016/j.coviro.2018.07.006

Okonechnikov, K., Golosova, O., and Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., et al. (2008). IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res.* 36, W163–W169. doi: 10.1093/nar/gkn198

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., et al. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. *BMC Genomic* 18:512. doi: 10.1186/s12864-017-3903-3

Pauvolid-Corrêa, A., Solberg, O., Couto-Lima, D., Nogueira, R. M., Langevin, S., and Komar, N. (2016). Novel viruses isolated from mosquitoes in pantanal, Brazil. *Genome Announc.* 4:e01195-16. doi: 10.1128/genomeA.01195-16

Pischedda, E., Scolari, F., Valerio, F., Carballar-Lejarazú, R., Catapano, P. L., Waterhouse, R. M., et al. (2019). Insights into an unexplored component of the mosquito repeatome: distribution and variability of viral sequences integrated into the genome of the arboviral vector *Aedes albopictus*. *Front. Genet.* 10:93. doi: 10.3389/fgene.2019.00093

Robinson, H. L., Astrin, S. M., Senior, A. M., and Salazar, F. H. (1981). Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. *J. Virol.* 40, 745–751. doi: 10.1128/jvi.40.3.745-751.1981

Ronquist, F., Huelsenbeck, J., and Teslenko, M. (2010). *Draft Mr. Bayes version 3.2 Manual*.

Russo, A. G., Kelly, A. G., Enosi Tuipulotu, D., Tanaka, M. M., and White, P. A. (2019). Novel insights into endogenous RNA viral elements in and other arbovirus vector genomes. *Virus Evol.* 5:vez010. doi: 10.1093/ve/vez010

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543. doi: 10.1038/nature20167

Sinha, A., and Johnson, W. E. (2017). Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene. *Curr. Opin. Virol.* 25, 105–112. doi: 10.1016/j.coviro.2017.07.020

Tasseto, M., Kunitomi, M., Whitfield, Z. J., et al. (2019). Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements. *eLife* 8:e41244. doi: 10.7554/eLife.41244

Théron, E., Dennis, C., Brasset, E., and Vaury, C. (2013). Distinct features of the piRNA pathway in somatic and germ cells: from piRNA cluster transcription to piRNA processing and amplification. *Mob. DNA* 5:28. doi: 10.1186/PREACCEPT-5823773114288275

Webb, B., and Sali, A. (2017). Protein structure modeling with modeller. *Funct. Genomics* 1654, 39–54. doi: 10.1007/978-1-4939-7231-9_4

Weiss, R. A. (2016). Exchange of genetic sequences between viruses and hosts. *Curr. Top. Microbiol. Immunol.* 407, 1–29. doi: 10.1007/82_2017_21

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., et al. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome. *Curr. Biol.* 27, 3511.e7–3519.e7. doi: 10.1016/j.cub.2017.09.067

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2011). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13-134

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.