bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Long-read assembly of the *Aedes aegypti* genome reveals the nature of heritable adaptive immunity sequences

Zachary J. Whitfield[1*], Patrick T. Dolan[1*], Mark Kunitomi[1*#], Michel Tassetto[1], Matthew G. Seetin[2], Steve Oh[2], Cheryl Heiner[2], Ellen Paxinos[2], and Raul Andino[1]

[1] Department of Microbiology and Immunology, University of California, 600 16th Street, GH-S572, UCSF Box 2280, San Francisco, California 94143-2280, USA.

[2] Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, California, 94025, USA.

*These authors contributed equally to this work

# Current address: IBM Almaden Research Center, 650 Harry Road, San Jose, California, 95120-6099, USA.

To whom correspondence should be addressed. E-mail: raul.andino@ucsf.edu

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The *Aedes aegypti* mosquito is a major vector for arboviruses including dengue, chikungunya and zika. Combating the spread of these viruses requires a more complete understanding of mosquito-virus interactions. Recent studies have implicated DNA derived from non-retroviral RNA viruses in insect immunity. To better define the role and origin of these elements, we generated a high-quality assembly of the *Ae. aegypti*-derived Aag2 cell line genome using single-molecule, real-time sequencing technology. The new assembly improves contiguity by one to two orders of magnitude with respect to previously released assemblies. This improved quality enables characterization of the collection of Endogenous Viral Elements (EVEs) in the mosquito genome, providing insight into their integration and role in mosquito immunity. Additionally, we find a distinct repertoire of EVEs present in the genomes of *Ae. aegypti* and *Ae. albopictus*, suggesting the intriguing possibility that differences in EVE composition may play a role in establishing vector competence.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Mosquito transmission of arboviruses such as Dengue virus (DENV), Chikungunya virus (CHIKV), and the newly emerging Zika virus causes widespread and debilitating disease across the globe (Bhatt et al., 2013). The primary vector of these viruses, *Aedes aegypti,* has a global tropical/subtropical distribution (Kraemer et al., 2015) creating geographically isolated populations of *Ae. aegypti* that have diversified over time. This genetic diversity has resulted in differential competence for vectoring virus(Bennett et al., 2002). Comparative genomics may explain these differences in vector competence between *Ae. aegypti* populations. However, this comparison is limited by the repetitive nature of the *Ae. aegypti* genome and the absence of an assembly based on long-read sequencing capable of spanning these regions. The underlying genetic contributions to vector competence are of critical importance to understanding the epidemiology of both acute epidemics as well as endemic occurrences of arbovirus infection.

One critical factor underlying vector competence is the insect immune system (Kramer, 2016; Kramer & Ciota, 2015). However, the connection between the mosquito immune system, tolerance to viral infection, and vector competence at a genomic level is still unclear. At their core, insects make use of an RNAi-based immune defense, where foreign viral dsRNA intermediates are recognized and processed through a dicer and argonaute mediated pathway, leading to cleavage of viral RNA and protection from infection (Mongelli & Saleh, 2016). In addition, mosquitoes utilize an additional RNAi pathway mediated by Piwi proteins and piRNAs as an antiviral defense system (Miesen, Joosten, & van Rij, 2016). Previous reports have indicated piRNAs can be produced from virus derived DNA sequences, and identified a set of proteins responsible for their processing and maturation (Goic et al., 2016; Miesen, Girardi, & van Rij, 2015; Miesen, Ivens, Buck, & van Rij, 2016; Miesen, Joosten, et al., 2016). Given that

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

endogenous viral elements (EVEs) (Katzourakis & Gifford, 2010) are capable of producing piRNAs (Arensburger, Hice, Wright, Craig, & Atkinson, 2011), we propose that rigorous EVE identification and examination in *Ae. aegypti* is central to understand arbovirus/vector interaction. However, due to the highly repetitive nature of the *Ae. aegypti* genome and EVEs' tendency to cluster within such repetitive regions, many EVEs are likely to be missing from the current *Ae. aegypti* assemblies, which are based on relatively short read lengths (Nene et al., 2007; Vicoso & Bachtrog, 2015).

Accordingly, we sought to generate a *de novo* genome assembly using long read deep sequencing, which is more suited to characterize the many repetitive regions that make up the majority of the *Ae. aegypti* genome. The resulting assembly greatly improves our understanding of the *Ae. aegypti* genome, a foundational tool for studying arboviral disease, spread, and prevention. As a demonstration of its utility, we use this improved assembly to characterize the structure and composition of EVE-containing loci across the entire *Ae. aegypti* genome.

Sequencing and Assembly of the Aag2 Genome

Current assemblies of the *Ae. aegypti* genome are based on two sequencing strategies: one produced with the Illumina sequencing platform (to be referred to as 'UCB') (Vicoso & Bachtrog, 2015) and one based on conventional Sanger sequencing (to be referred to as 'LVP') (Nene, et al., 2007). In all instances, the Liverpool strain of *Ae. aegypti* was sequenced (Table 1). A more recent study used Hi-C to further organize the sanger-based *Ae. aegypti* assembly into chromosome level scaffolds (Dudchenko et al., 2017). Long-read assembly will further aid these efforts by reading through and resolving large repetitive regions unable to be identified by previous methods. To this end, we employed single-molecule, real-time sequence technology (Pacific Biosystems) to generate long read sequences of the genome of the cell line Aag2. Of note, the *Ae. aegypti*-derived Aag2 cell line is an important, widely used tool in the study of *Ae. aegypti* biology and its associated arboviruses. Here, we present the

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

draft assembly of the Aag2 genome and use it to map endogenous viral elements (EVEs) genome-wide.

Approximately 76-fold coverage of the *Ae. aegypti*-based Aag2 genome was achieved using the Single Molecule Real Time (SMRT) sequencing platform (P6/C4 chemistry) to shotgun sequence 116 SMRT cells generating 92.7 GB of sequencing data with an average read length of 15.5 kb. We used Falcon and Quiver to generate a *de novo* 1.7 Gbp assembly with a contig N50 of ~1.4 Mbp.

Our assembly improves upon previous *Aedes* assemblies as measured by N50, L50, and simply by the number of contigs (Table 1 and Figure 1). The long-read sequencing approach enabled a substantial improvement in assembly contiguity (Figure 1a). A majority of the Aag2 assembly sequence is found on contigs 10-100x longer than previous assemblies. This increased contiguity allows the mapping of numerous contigs from the initial LVP assembly to single Aag2 contigs (Figure 1b and c), and makes for an overall more complete and much more ordered genome assembly for use in downstream processes. In this report, we focused on examining repetitive regions of the genome, and as expected, uncovered a plethora of Endogenous Viral Elements (EVEs). These EVEs were found to produce *bona fide* small RNAs, and were organized into loci akin to a CRISPR-like system(de Vanssay et al., 2012).

Repetitive nature of the Aag2 genome

The genome of *Ae. aegypti* was previously shown to contain high proportions of repeat-DNA (Nene, et al., 2007). The *Ae. aegypti*-derived Aag2 genome is no different, and is comprised of almost 55% repeat sequence (Table 1 and Table 2). Our sequencing strategy allows more repeats to be sequenced within a single read, and therefore better reflects the structure and organization of these repetitive elements. Direct alignment of contigs in the Aag2 assembly and those of previous *Ae. aegypti* assemblies reveal resolved rearrangements and distinct repeated regions that were collapsed into single sequences in the

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

previous assemblies (Figure 1c and d). These regions can span 10-20kb (uncollapsed), illustrating the need for long read lengths to properly order the vast amount of repetitive regions in the *Ae. aegypti* genome. Of these repetitive regions, over 75% is made up of transposon-derived sequence (Table 2).

Transposon-derived sequences in the *Ae. aegypti* genome

Transposable elements (TEs) play important roles in gene regulation and genome evolution, providing a source of genomic variation that is a driver of evolution (Gifford, Pfaff, & Macfarlan, 2013; Thompson, Macfarlan, & Lorincz, 2016). TEs account for a large portion of the *Aedes aegypti* genome and our long-read assembly allows us to explore the large-scale structure of these prominent features (Figure 2a and b). Individual contigs of the Aag2 assembly contained on average a significantly higher number of transposons than previously observed in the other two assemblies (Figure 2c).

The TEs in the Aag2 genome are derived from a number of different families and distributed throughout the genome (Figure 2a, d , Figure 2-Figure Supplement 1). However, the distribution of TEs is not completely uniform across the genome (Figure 2e). Local density plots reveal regions where particular TE classes are overrepresented, likely reflecting integration site bias and/or piRNA cluster formation (discussed below). Kimura distribution analysis (Kimura, 1980) of TEs in the Aag2 genome shows a relatively recent expansion of TEs, particularly LINE, LTR, and MITEs elements (Figure 2f; low Kimura scores indicate TEs that are closer to the element's consensus sequence, while higher scores indicated more diverged TE sequences). With our assembly a resource particularly suited for repeat identification and analysis has been produced.

Identification of EVEs

Given their propensity to integrate into repetitive TE clusters (Figure 3b)(Parrish et al., 2015), our understanding of the EVE composition and structure has been limited. We thus used our improved long-read assembly genome to

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

better define the complete set of EVEs contained within the Aag2 genome, hereby called the "EVEome". Using a BLASTx-based approach (see Methods), we searched for all EVE sequences within the newly assembled Aag2 genome. A total of 368 EVEs were identified in the Aag2 assembly (Figure 3a). We were able to detect EVEs derived from at least 8 viral families, dominated by sequences derived from Rhabdoviridae, Flaviviridae, and Chuviridae (Figure 3c).

Given that TE clusters have been shown to produce piRNAs, we next examined the populations of these small RNAs mapping to EVEs across the entire Aag2 genome (Figure 3a, b, and d). While only a small proportion of *Ae. aegypti* TE elements produce piRNAs(Arensburger, et al., 2011), we observed 281 EVEs (73.2% of all EVEs) each producing at least 10 (Figure 3a, red lines). Characterization of these EVE piRNAs suggest they are *bona fide* piRNAs because they are methylated (resistant to β–elimination), present a 5' U bias, and range in size from 24 to 30 nucleotides) (data not shown). Furthermore, when compared to the previous *Ae. aegypti* assemblies, ~2x10^6 more piRNAs could be mapped to the more contiguous Aag2 assembly (Figure 3d), indicating our assembly gives a more complete tool for the study of the EVE composition within repetitive regions of the genome. With their propensity to produce mature piRNAs, and their integration in the *Ae. aegypti* genome, we speculate that EVEs comprising the "EVEome" represent a collection of adaptive immune "cassettes" in *Ae. aegypti,* some of which could function anti-virally.

Insights into the mechanism of EVE integration

DNA derived from RNA viruses is produced in persistently infected Drosophila cell lines (Goic et al., 2013) and in infected *Aedes albopictus* mosquitoes (and multiple mosquito-derived cell culture lines) (Goic, et al., 2016). This viral DNA (vDNA) synthesis depends on the activity of endogenous reverse transcriptases (Goic, et al., 2016; Goic, et al., 2013). Further, sequencing of viral DNA isolated from Drosophila cell lines (Goic, et al., 2013) has demonstrated formation of DNA hybrids between viral DNA sequences and transposable

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

elements. EVEs are typically found nearby transposons and can be found within piClusters (Figure 3a and b) (Feschotte & Gilbert, 2012; Honda & Tomonaga, 2016; Miesen, Joosten, et al., 2016). Together this information suggests that EVEs are generated and integrated into the host genome in the context of the replication cycle of endogenous transposable elements. To determine whether a particular transposable element type is responsible for EVE integration, we identified TEs whose coordinates in the genome directly overlap EVE sequences (as called by RepeatMasker and BLASTX respectively). This approach identifies mobile elements most likely responsible for genomic integration of non-retroviral virus sequence. In line with observations in *Ae. albopictus (X. G. Chen et al., 2015)*, LTR retrotransposons were found to be greatly enriched near EVEs (Figure 4-Figure Supplement 1 (i)). A similar pattern was observed when classifying the nearest upstream and downstream (non-overlapping) TE sequences around each EVE (Figure 4a(i)). These results further implicate LTR retrotransposons in the acquisition of EVEs and indicate that the typical integration sites are composed of clusters of similar LTR retrotransposons.

Strikingly, a vast majority of these LTR TEs shared the same strandedness as their nearest EVE. Of 571 (non-overlapping) TEs with the same strandedness as their nearest EVE, 419 were LTRs (Figure 4a(i); p-value = 3.42×10^{-188} by one-sided binomial test). This bias is consistent with a copy-choice mechanism of recombination between LTR retrotransposon sequence and viral RNA leading to EVE integration, as previously proposed (Cotton, Steinbiss, Yokoi, Tsai, & Kikuchi, 2016; Geuking et al., 2009). Our analysis of transposons in the Aag2 genome shows LTR-retrotransposons display less diversity (by Kimura Divergence score; Figure 2f), indicating that they are currently (or were recently) actively replicating in the Aag2 cell line. Consistent with this idea, LTR-retrotransposon transcripts and proteins are readily detected in Aag2 cells (Maringer et al., 2017). We thus conclude that LTR-retrotransposons are responsible for the acquisition of the majority of EVEs observed in the *Ae. aegypti* genome.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Within the LTR retrotransposon family, both Ty3/gypsy and Pao Bel TEs are enriched surrounding EVEs (Figure 4a iv,v). Again, this enrichment for Ty3/gypsy and Pao Bel elements near EVE loci is strongest when the EVE and TE are in the same orientation (p-value = 7.99×10^{-23} and 2.00×10^{-3} respectively). The drastic reduction in associated transposons based on directionality is not observed for other TE categories (Figure 4a, iii, vi). These data support Ty3/gypsy (and to a lesser extent Pao Bel) as the primary transposon type facilitating EVE genomic integration in *Ae. aegypti*. Interestingly, an association between LTR Ty3/gypsy elements and integrated viral sequence has also been observed previously in plants (Lee, Nolan, Watson, & Tristem, 2013; Staginnus et al., 2007), suggesting a conserved mechanism for the acquisition of invading virus sequences and generation of EVEs.

EVE-proximal TEs of the Ty3/gypsy and Pao Bel families can be further partitioned into individual elements. Of these, many specific elements were enriched for being the nearest TE to an EVE (Figure 4a(iv, v)). Interestingly, EVEs derived from different virus families show different patterns of enrichment for nearby TEs (Figure 4b). Both Flaviviridae and Rhabdoviridae-derived EVEs show strong enrichment for Ty3/gypsy transposable elements, while Chuviridae-derived EVEs are most typically adjacent to Pao Bel elements. Of the Ty3 elements near Rhabdoviridae and Flaviviridae-derived EVEs, Ele152 is most enriched near Flaviviridae sequences, while Ty3/gypsy Ele134 and Ele135 are most closely associated with Rhabdoviridae EVE sequences. We hypothesize these data reflect the particular mobile elements whose replication overlapped with a given RNA virus' replication (in both space and time).

The strong enrichment for multiple LTRs around EVEs (Figure 3b, 4a, Figure 4-Figure Supplement 1), as well as the presence of EVE-derived piRNAs (Figure 3a), are consistent with previous observations that EVEs integrate into

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

piRNA clusters (Parrish, et al., 2015). We identified 469 piRNA-encoding loci (piClusters) using proTRAC (Rosenkranz, Rudloff, Bastuck, Ketting, & Zischler, 2015; Rosenkranz & Zischler, 2012), accounting for 5,774,304 bp (0.335%) of the genome. Depending on the mapping algorithm used, between 63% (bowtie) and 77% (sRNAmapper.pl, see Methods) of beta-eliminated small RNAs from Aag2 cells mapped to these loci. Of the identified piClusters, 66 (14.1%) have EVE sequences associated with them and 65 of these piCluster-resident EVE sequences act as the template for piRNAs. Of the 384 EVEs identified, 256 (66.7%) or 280,475 bp of the 411,239 EVE bp mapped to piClusters (68.2%, Fisher's test p<2.2e-16, OR=203.42). Furthermore, a vast majority of piRNAs which map to EVEs are anti-sense to the EVE itself (544,429/547,014; 99.5%), consistent with the idea that EVEs produce functional piRNAs and are selected through evolution for their antiviral potential.

EVE/TE piRNA clusters are typically found in unidirectional orientations (Figure 3b, Figure 3-Figure Supplement 1), and strikingly, some LTR/EVE clusters have 'crowded out' almost all other transposons in the same region (Figure 3b). We hypothesize these TE/EVE clusters are the result of canonical piRNA cluster formation (Yamanaka, Siomi, & Siomi, 2014), with the rare occurrence of LTR-facilitated EVE integration, resulting in localized genomic expansion (due to non-random LTR integrase directed integration)(Lesbats, Engelman, & Cherepanov, 2016). These data together support the notion that EVEs are located at specific genomic loci associated with high piRNA production.

EVEs in *Ae. aegypti* and *Aedes albopictus*

If EVEs serve as a representative record of viral infection over time, and the majority are a 'recent' acquisition (based on Ty3/gypsy's low Kimura Divergence scores (Figure 2-Figure Supplement 2)), we hypothesized that EVEs present in two different species of mosquito would also differ (particularly given the relatively rare occurrence of genome fixation) (Holmes, 2011; Katzourakis & Gifford, 2010). The *Ae. aegypti* and *Ae. albopictus* species of mosquito occupy

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distinct (yet overlapping) regions around the globe (Kraemer, et al., 2015) and so have not faced the exact same viral challenges over time. While the EVEs present in the Aag2 and LVP *Ae. aegypti*-based genomes correspond well, *Ae. aegypti* and *Ae. albopictus* do not share any specific EVEs. However, exploring the Flaviviridae family of viruses in greater detail, the viral species from which these EVEs are derived do partially overlap (Figure 5a). However, the relative abundance of EVEs derived from various viral species in *Ae. aegypti* and *Ae. albopictus* differs. The lack of specific EVEs in common between the mosquito genomes indicates EVE acquisition by *Ae. aegypti and Ae. albopictus* occurred post-speciation, an important factor when considering any differences in vector competence between these two species.

We next determined whether any particular region of the virus genome is more frequently acquired and converted into EVEs. EVEs were mapped to the locations of the viral ORFs from which they originate (Figure 5b). Interestingly, EVEs deriving from Flaviviridae primarily map toward the 5' end of the single Flaviviral ORF, leaving a relative dearth of EVEs at the 3' end. EVEs deriving from Rhabdoviridae primarily originate from the Nucleoprotein (N) and Glycoprotein (G) coding sequences, with only a few originating from the RNA-dependent RNA polymerase (L). The lack of EVEs mapping to the polymerase may be the result of RNA expression levels, with L being the least expressed gene (Conzelmann, 1998). This suggests that the template for cDNA synthesis and recombination with a TE genome are viral mRNAs. The lack of EVEs originating from the Phosphoprotein (P) or Matrix protein (M) is more difficult to explain, potentially reflecting the localization and/or availability of the RNA species for recombination. Interestingly, EVEs derived from Chuviridae primarily map to the ORF of the Glycoprotein. Given the complex and diverse nature of Chuviridae genomes (which include unsegmented, bi-segmented, and possibly circularized negative-sense genomes)(Li et al., 2015), this pattern could also be the result of RNA abundance. Within each ORF however, there is no obvious

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

selection for EVEs from a particular location. Rather, EVEs map to regions evenly distributed across their respective ORFs.

We next analyzed the Kimura divergence scores for TEs closest to (but not overlapping) EVEs (both upstream and downstream; Figure 5c). When TEs were grouped both by family and the family of the virus from which its closest EVE was derived from, stark differences in Kimura score distribution emerge (Figure 5c, Figure 5-Figure Supplement 1). Pao Bel elements near EVEs originating from Chuviridae and Flaviviridae had much higher Kimura divergence scores than Ty3/gypsy elements matching the same criteria. If Kimura divergence score is taken as a proxy for ‘genomic age’, these Pao Bel elements (and potentially the EVEs they are most closely associated with) are a more ancient acquisition than their Ty3/gypsy counterparts. On the other hand, both Pao Bel and Ty3/gypsy elements closest to EVEs derived from Rhabdoviridae show a more uniform distribution of scores, suggesting there may have been continued periods of Rhabodviral infection (and subsequent EVE acquisition by transposable elements) in *Ae. aegypti*.

Recent publications have highlighted the integration of genetic material from non-retroviral RNA viruses into the genome of the host during infection that relies upon endogenous retro-transcriptase activity from transposons. A subset of EVEs found in mammalian systems seem to be under purifying selection, which suggests that they are beneficial to the host (Horie et al., 2010). A species’ “EVEome” could represent a built-in, yet adaptable, viral defense system. All mosquito species share the same basic RNAi-based immune system. We propose differences in a given specie’s (or subpopulation’s) “EVEome”, such as those between *Ae. aegypti* and *Ae. albopictus* (Figure 5a), may represent an important factor contributing to inherent differences in vector competence. In support of this hypothesis, the *Ae. albopictus* assembly contains an EVE that clusters phylogenetically with extant DENV (Figure 5a) and has been shown to

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be less competent at disseminating DENV in some cases (W. J. Chen, Wei, Hsu, & Chen, 1993; Whitehorn et al., 2015).

Discussion

In vivo studies of mosquito immunity are a valuable, but challenging approach to understanding arboviral life cycles. The *Ae. aegypti*-based cell line Aag2 provides a tool with which molecular characterization of the arboviral replication in mosquitos can be accomplished on a much wider scale. An improved, fully assembled *Ae. aegypti*-derived Aag2 genome is a significant step in best utilizing this cell line and furthering our understanding of *Ae. aegypti* biology.

The presence of piRNA producing EVEs in the *Ae. aegypti* genome is reminiscent of the CRISPR system in bacteria. Both take advantage of the invading pathogen's genetic material to create small RNAs capable of restricting an invading virus' replication. Furthermore, both end up integrated into the host's genome, likely providing some level of protection against future infections. EVEs themselves can present a unique opportunity to track viral evolution and historical interactions between host and virus (Holmes, 2011; Katzourakis, Tristem, Pybus, & Gifford, 2007; Keckesova, Ylinen, Towers, Gifford, & Katzourakis, 2009). Our Aag2 genome assembly refines our understanding of EVEs in the *Ae. aegypti* genome, and the breadth of coverage by which they may protect hosts against viral infections.

The finding that LTR retrotransposon elements (specifically Ty3/gypsy and Pao Bel) are closest in proximity to EVEs is intriguing. A number of (non-mutually exclusive) possibilities could explain this observation. It may be that Ty3/gypsy and Pao Bel elements were in the 'right place at the right time' to participate in acquisition of viral sequences as EVEs. LTR retrotransposon replication and virus replication must have overlapped both physically within the cell and

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

temporally within the natural history of *Ae. aegypti* evolution so as to provide the opportunity for LTR-mediated EVE integration. Template switching during reverse transcription has previously been proposed to play a role in creating the transposon-virus hybrids which integrate into the host genome to form EVEs (Cotton, et al., 2016; Geuking, et al., 2009). The enrichment of Pao Bel TEs near Chuviridae-derived sequences and Ty3/gypsy TEs near Flaviviridae and Rhabdoviridae sequences (Figure 4b) could have occurred by chance, or may hint at an even deeper level of specificity directing capture of viral sequences by LTRs. Possibly these TE/EVE pairs have increased sequence homology leading to more frequent template switching of the reverse transcriptase (Delviks-Frankenberry et al., 2011), or their subcellular localization of replication better coincide. One key distinction between LTR and non-LTR retrotransposons is the cellular location in which reverse transcription (RT) occurs. LTR retrotransposons undergo reverse transcription in the cytoplasm, while non-LTR retrotransposons undergo RT within the nucleus (Servant & Deininger, 2015). Given most RNA viruses also replicate within the cytoplasm, the opportunity for template switching onto an RNA virus genome would be more readily available to an LTR retrotransposon.

It is also possible that only LTR/EVE pairs were selected for after integrating into the mosquito genome. The role of selection on the organization and placement of EVEs in the *Ae. aegypti* genome is an intriguing one, and many EVEs have integrated into piRNA clusters from which *bona fide* piRNAs are produced (Figure 3a,b). However, a lack of synteny between EVEs in *Ae. aegypti* and *Ae. albopictus* suggests their acquisition likely took place post-speciation. Given this, along with LTRs being most active most recently (among TEs) in the Aag2 cell line (Figure 2f) (Maringer, et al., 2017), many EVEs in the *Aedes aegypti* genome likely represent relatively recent acquisitions. While an informative observation, this also limits the ability to date these EVEs with molecular-clock based techniques, and thus properly study selection. However,

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the evidence so far suggests that some EVE sequences (and their organization in the genome) are being maintained evolutionarily.

As mentioned above, only the *Ae. albopictus* genome contains EVEs which cluster with extant DENV, and it also happens to be a less suitable vector for the virus (as compared to *Ae. aegypti*). We hypothesize these particular EVE insertions may provide a buffer against DENV infection. *Ae. albopictus* is still susceptible to DENV infection, thus the DENV-derived EVE does not provide complete protection against DENV infection. However, as piRNAs are found at various levels within different tissues of the mosquito (Akbari et al., 2013), EVE-derived piRNAs likely confer differing levels of viral resistance depending on the tissue in question. Thus, EVE-derived piRNAs could play a role in keeping viral infections in check at an organismal level, maybe indefinitely in the case of persistent infections. The hypothesis that many EVE-derived piRNAs in *Ae. aegypti* are functional (and their organization in the genome has been preserved) is further supported by the shared orientation of LTR retrotransposons and their nearby EVEs. In this arrangement, precursor transcripts originating from EVE/TE-containing piRNA clusters encode anti-viral piRNA sequence with the same directionality they contain anti-TE piRNA sequence. The resulting anti-sense piRNAs can then go on to silence their complementary (viral or TE) RNA.

A solid foundation with which to study the genetic factors contributing to vector competence is of utmost importance. With this in mind, we generated a highly contiguous assembly of the *Aedes aegypti* cell line, Aag2. With this long-read assembly, we then identified nearly the entire set of endogenous viral elements and their surrounding genomic context in the Aag2 cell line at a genome-wide scale. Uncovering the genomic context of this EVE-derived piRNA system in mosquitos provides the foundation for future studies on the role of EVEs in vector competence. The potential to manipulate a heritable, anti-viral system opens up new avenues to understand the complexities of the insect immune system and work to prevent spread of viral disease dependent on such

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

insect vectors.

Acknowledgements

We thank Dr. Kevin M. Dalton for helpful discussion and code for the analysis.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MATERIALS AND METHODS

Cells culture

Aedes aegypti Aag2 *(Lan & Fallon, 1990; Peleg, 1968)* cells were cultured at 28 °C without CO_2 in Schneider's *Drosophila* medium (GIBCO-Invitrogen), supplemented with 10% heat-inactivated fetal bovine serum (FBS), 1X non-essential amino acids (NEAA, UCSF Cell Culture Facility, 100X stock is 0.1 µM filtered, 10 mM each of Glycine, L-Alanine, L-Asparagine, L-Aspartic acid, L-Glutamic Acid, L-Proline, and L-Serine in de-ionized water), and 1X Penicillin-Streptomycin-Glutamine (Pen/Strep/G, 100X = 10,000 units of penicillin, 10,000 µg of streptomycin, and 29.2 mg/ml of L-glutamine, Gibco).

DNA sequencing

Aag2 cells were grown in T-150 Flasks until ~80% confluent. Cells were then washed with dPBS twice and scrapped off in dPBS + 10 µg/ml RNase A (ThermoFisher). Genomic DNA (gDNA) was extracted from ~10^8 Aag2 cells using the QIAamp DNA Mini Kit according to the manufacturer's instructions with the optional RNase A treatment. Aag2 gDNA was re-suspended in 10mM Tris pH8, and the quality and quantity of the sample was assessed using the Agilent DNA12000 kit and 2100 Bioanalyzer system (Agilent Technologies), as well as the Qubit dsDNA Broad Range assay kit and Qubit Fluorometer (Thermo Fisher) and visualized by gel electrophoresis (1% TBE gel). After purification and quality control, a total of 130 ug of DNA was available for library preparation and sequencing.

SMRTbell libraries were prepared using Pacific Biosciences' Template Prep Kit 1.0 (PacBio) and a slightly modified version of the Pacific Biosciences' protocol, "Procedure & Checklist - 20-kb Template Preparation Using BluePippin Size-Selection System (15-kb Size Cutoff)". Specifically, 52.5ug of gDNA were hydrodynamically sheared to target sizes of 30kb (26 µg) and 35 kb (26 µg) using the Megaruptor® (Diogenode) with long hydropores according to the manufacturer's protocols. Size distributions of the final sheared gDNA were verified by pulse field electrophoresis of a 100ng sub-aliquot through 0.75%

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

agarose using the Pippin Pulse (Sage Science), run according to the manufacturer's "10-48 kb protocol" for 16 hrs. The two sheared samples were then pooled, for a total of 37ug sheared DNA to be used as input into SMRTbell preparation. Sheared DNA was subjected to DNA damage repair and ligated to SMRTbell adapters. Following ligation, extraneous DNA was digested with exo-nucleases and the resulting SMRTbell library was cleaned and concentrated with AMPure PB beads (Pacific Biosciences). A total of 20.5ug of library was available for size selection.

Approximately half (10ug) of the SMRTbell pooled SMRTbell library was size-selected using the BluePippin System (Sage Science) using a 15 kb cutoff and 0.75% agarose cassettes. To obtain longer read lengths, an additional 5ug of the library was selected using a 17kb cutoff.

Library quality and quantity were assessed using the Agilent 12000 DNA Kit and 2100 Bioanalyzer System (Agilent Technologies), as well as the Qubit dsDNA Broad Range Assay kit and Qubit Fluorometer (Thermo Fisher). An additional DNA Damage Repair step and AMPure bead cleanup were included after size-selection of the libraries.

Annealed libraries were then bound to DNA polymerases using 3nM of the SMRTbell library and 3X excess DNA polymerase at a concentration of 9nM using Pacific Biosciences DNA/Polymerase Binding Kit 1.0, (Pacific Biosciences). Bound libraries were sequenced on the Pacific Biosciences RSII using P6/C4 chemistry (PacBio), magnetic bead loading (PacBio) and 6 hour collection times. 84 SMRTcells of the > 15 kb library were loaded at concentrations of 75-100 pM on-plate. 32 SMRTcells of the > 17 kb library was prepared separately and loaded at on-plate concentrations of 40 pM and 60 pM. These 116 SMRTcells generated 92.7 GB of sequencing data, which resulted in approximately 76X coverage of the Aag2 genome. Average raw read length of 15.5KB, with average sub-reads length of 13.2kb. Assembly was performed using Quiver/FALCON

Genome assembly statistics

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Basic statistics (e.g. Size, Gaps, N50, L50, # contigs) for each genome analyzed was produced using Quast (Gurevich, Saveliev, Vyahhi, & Tesler, 2013).

As a complementary approach Benchmarking sets of Universal Single-Copy Orthologs (BUSCO) was also run using the Arthropod dataset in order to assess the completeness of genome assembly. Of the 2675 BUSCO groups searched only 81 were missing from the Aag2 assembly, indicating good assembly completeness. Of the 2315 BUSCOs found only 279 of them were annotated as fragmented, emphasizing the continuity of the assembly.

Repeat Identification and Kimura Divergence

In order to *de novo* identify and classify novel repetitive elements from the Aag2 genome, RepeatModeler was run on the assembled genome using standard parameters. Outputs from RepeatModeler were cross-referenced with annotated entries for Aedes aegypti from TEfam. All entries from RepeatModeler that were >80% identical to TEfam entries were discarded as redundant. This combined annotated and de novo identified list of repeat elements was used to identify the genome wide occurrences of repeats using RepeatMasker using standard parameters.

Kimura scores and corresponding alignment information were extracted from the ".align" file as output by RepeatMasker. This information was used to make the stacked plot in figure 2 using R (version 3.30) and the ggplot2 package.

Information from the header lines was then used to match Kimura divergence score with the appropriate EVE-proximal TEs based on contig, TE name, start point, and end point. TEs whose coordinates did not exactly match output of the align file were not used in the Kimura analysis. The violin plot of Kimura divergence scores was plotted using R (version 3.3.0) and the ggplot2 plugin.

Citations:

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*.
2008-2015 <http://www.repeatmasker.org>
Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*.
2013-2015 <http://www.repeatmasker.org>.
Dr. Zhijian Jake Tu. TEfam. <http://tefam.biochem.vt.edu/tefam/index.php>

EVE identification

Identification of EVEs was achieved using standalone Blast+ (Altschul, Gish, Miller, Myers, & Lipman, 1990). Blast Searches were run using the Blastx command specifying the genome as the query and a refseq library composed of the ssRNA and dsRNA viral protein-coding sequences from the NCBI genomes as the database. The E-value threshold was set at 10-6.

The EVE with the lower E-value was chosen for further analysis to predict EVEs that overlapped. Several Blast hits to viral protein genes were identified as artifacts because of their homology to eukaryotic genes (e.g. closteroviruses encode an Hsp70 homologue). These artifacts were filtered by hand.

Identification of LTR enrichment near EVEs

Separate BED files containing all TEs in the Aag2 assembly and all EVEs in the Aag2 assembly were used as input to Bedtools (*bedtools closest* command using the –*io* flag, and –*id* or -*iu*) to find the single closest non-overlapping TE to each EVE (both upstream and downstream).

An in-house script compiled these two output files together and filtered them for the TE content of interest. TE categories (subclass, family, element) were assigned by RepeatMasker. Enrichment was compared to the prevalence of the TE element genome wide based on a one-sided binomial test. Stacked histograms were produced based on TE categories as found in Figure 3. The legend lists (up to) the 10 most prevalent TE elements of TE/EVE pairs in the same orientation. Plots were produced using Python (version 2.7.6) with the pandas and matplotlib plugins.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Classification of nearest TE to EVEs by virus taxonomy

Taxonomy categories for viruses from which each EVEs derived were assigned using an in-house script. Assignments were made based on NCBI's taxonomy database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/), with the following additional annotations by hand.

Virus species; assigned family:

Wuhan Mosquito Virus 8; Chuviridae
Wuchang Cockraoch Virus 3; Chuviridae
Lishi Spider Virus 1; Chuviridae
Shayang Fly Virus 1; Chuviridae
Wenzhou Crab Virus 2; Chuviridae

Bole Tick Virus 2; Rhabdoviridae
Shayang Fly Virus 2; Rhabdoviridae
Wuhan Ant Virus; Rhabdoviridae
Wuhan Fly Virus 2; Rhabdoviridae
Wuhan House Fly Virus 1; Rhabdoviridae
Wuhan Mosquito Virus 9; Rhabdoviridae
Yongjia Tick Virus 2; Rhabdoviridae

Cilv-C; Virgaviridae
Citrus leprosis virus C; Virgaviridae
Blueberry necrotic ring blotch virus; Virgaviridae

Wutai Mosquito Virus; Bunyaviridae

Heat maps were produced using the Seaborn plugin for python. Only TEs with >=10% proportion in at least one sample (Flaviviridae, Chuviridae, or Rhabdoviridae) are shown. Color was assigned based on proportion of TE element/family in each viral category.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Enrichment was scored as above using a one-sided binomial test (significant is p-value < 0.0001).

Small RNA bioinformatics

Adaptors were trimmed using Cutadapt (http://dx.doi.org/10.14806/ej.17.1.200) using the --discard-untrimmed and -m 19 flags to discard reads without adaptors and below 19 nt in length. Reads were mapped using bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009)using the –v 1 flag. Read distance overlaps were generated by viROME (Watson, Schnettler, & Kohl, 2013). Sequence biases were determined by Weblogo (Crooks, Hon, Chandonia, & Brenner, 2004).

piCluster Analysis

piClusters were identified using PROtrac (Rosenkranz & Zischler, 2012) based on mapping with positions for beta-eliminated small RNAs libraries from Aag2 cells from sRNAmapper.pl. Based on these predictions, visualizations of clusters were produced using EasyFig (Sullivan, Petty, & Beatson, 2011) for visualization of TEs and R for comparison of TEs, piRNA abundance and EVE positions.

Sequence alignment and phylogenetic analysis

For phylogenetic analysis of Flaviviridae, polyprotein sequences from 61 members of the Flaviviridae family were aligned with MUSCLE (Edgar, 2004) and a maximum likelihood tree was generated with FastTree (Price, Dehal, & Arkin, 2009) using the generalised time reversible substitution model (“-gtr”). Trees were visualized and annotated with ggtree (DOI: 10.1111/2041-210X.12628).

EVE coverage

Base R (version 3.3.0) was used to show regions individual EVEs span on the indicated viral family (and protein). EVE length is a function of the percentage of the respective ORF from which it derives.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Statistics summary

Enrichment for TE elements near EVEs (Figure 3B, 3F) was determined with a one-sided binomial test (alternative hypothesis ‘greater’). Enrichment of EVEs in piClusters was determined by Fisher’s Test. Difference in Kimura Divergence distributions (not necessarily normally distributed) of TEs near EVEs vs total EVE populations (Figure 4C) was determined by Kolmogorov-Smirnov tests.

Code availability

The code used to generate the datasets used for visualization have been provided.

Data availability

The Aag2 genome (v 1.00) is available through VectorBase (https://www.vectorbase.org/organisms/aedes-aegypti/aag2/aag2). Main datasets produced during this work have been provided in excel format.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFERENCES

Akbari, O. S., Antoshechkin, I., Amrhein, H., Williams, B., Diloreto, R., Sandler, J., & Hay, B. A. (2013). The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. *G3 (Bethesda), 3*(9), 1493-1509. doi: 10.1534/g3.113.006742

g3.113.006742 [pii]

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

S0022-2836(05)80360-2 [pii]

Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L., & Atkinson, P. W. (2011). The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics, 12*, 606. doi: 10.1186/1471-2164-12-606

1471-2164-12-606 [pii]

Bennett, K. E., Olson, K. E., Munoz Mde, L., Fernandez-Salas, I., Farfan-Ale, J. A., Higgs, S., . . . Beaty, B. J. (2002). Variation in vector competence for dengue 2 virus among 24 collections of Aedes aegypti from Mexico and the United States. *Am J Trop Med Hyg, 67*(1), 85-92.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., . . . Hay, S. I. (2013). The global distribution and burden of dengue. *Nature, 496*(7446), 504-507. doi: 10.1038/nature12060

nature12060 [pii]

Chen, W. J., Wei, H. L., Hsu, E. L., & Chen, E. R. (1993). Vector competence of Aedes albopictus and Ae. aegypti (Diptera: Culicidae) to dengue 1 virus on Taiwan: development of the virus in orally and parenterally infected mosquitoes. *J Med Entomol, 30*(3), 524-530.

Chen, X. G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., . . . James, A. A. (2015). Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. *Proc Natl Acad Sci U S A, 112*(44), E5907-5915. doi: 10.1073/pnas.1516410112

1516410112 [pii]

Conzelmann, K. K. (1998). Nonsegmented negative-strand RNA viruses: genetics and manipulation of viral genomes. *Annu Rev Genet, 32*, 123-162. doi: 10.1146/annurev.genet.32.1.123

Cotton, J. A., Steinbiss, S., Yokoi, T., Tsai, I. J., & Kikuchi, T. (2016). An expressed, endogenous Nodavirus-like element captured by a retrotransposon in the genome of the plant parasitic nematode Bursaphelenchus xylophilus. *Sci Rep, 6*, 39749. doi: 10.1038/srep39749

srep39749 [pii]

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190. doi: 10.1101/gr.849004

14/6/1188 [pii]

de Vanssay, A., Bouge, A. L., Boivin, A., Hermant, C., Teysset, L., Delmarre, V., . . . Ronsseray, S. (2012). Paramutation in Drosophila linked to emergence of a piRNA-producing locus. *Nature, 490*(7418), 112-115. doi: 10.1038/nature11416

nature11416 [pii]

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Delviks-Frankenberry, K., Galli, A., Nikolaitchik, O., Mens, H., Pathak, V. K., & Hu, W. S. (2011). Mechanisms and factors that influence high frequency retroviral recombination. *Viruses, 3*(9), 1650-1680. doi: 10.3390/v3091650
viruses-03-01650 [pii]

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. doi: eaal3327 [pii]
10.1126/science.aal3327
science.aal3327 [pii]

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi: 10.1093/nar/gkh340
32/5/1792 [pii]

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet, 13*(4), 283-296. doi: 10.1038/nrg3199
nrg3199 [pii]

Geuking, M. B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., . . . Hangartner, L. (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science, 323*(5912), 393-396. doi: 10.1126/science.1167375
323/5912/393 [pii]

Gifford, W. D., Pfaff, S. L., & Macfarlan, T. S. (2013). Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol, 23*(5), 218-226. doi: 10.1016/j.tcb.2013.01.001
S0962-8924(13)00003-2 [pii]

Goic, B., Stapleford, K. A., Frangeul, L., Doucet, A. J., Gausson, V., Blanc, H., . . . Saleh, M. C. (2016). Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat Commun, 7*, 12410. doi: 10.1038/ncomms12410
ncomms12410 [pii]

Goic, B., Vodovar, N., Mondotte, J. A., Monot, C., Frangeul, L., Blanc, H., . . . Saleh, M. C. (2013). RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. *Nat Immunol, 14*(4), 396-403. doi: 10.1038/ni.2542
ni.2542 [pii]

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075. doi: 10.1093/bioinformatics/btt086
btt086 [pii]

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe, 10*(4), 368-377. doi: 10.1016/j.chom.2011.09.002
S1931-3128(11)00285-X [pii]

Honda, T., & Tomonaga, K. (2016). Endogenous non-retroviral RNA virus elements evidence a novel type of antiviral immunity. *Mob Genet Elements, 6*(3), e1165785. doi: 10.1080/2159256X.2016.1165785
1165785 [pii]

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., . . . Tomonaga, K. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature, 463*(7277), 84-87. doi: 10.1038/nature08695
nature08695 [pii]

Katzourakis, A., & Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet, 6*(11), e1001191. doi: 10.1371/journal.pgen.1001191

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Katzourakis, A., Tristem, M., Pybus, O. G., & Gifford, R. J. (2007). Discovery and analysis of the first endogenous lentivirus. *Proc Natl Acad Sci U S A, 104*(15), 6261-6265. doi: 0700471104 [pii]
10.1073/pnas.0700471104

Keckesova, Z., Ylinen, L. M., Towers, G. J., Gifford, R. J., & Katzourakis, A. (2009). Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology, 384*(1), 7-11. doi: 10.1016/j.virol.2008.10.045
S0042-6822(08)00718-6 [pii]

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol, 16*(2), 111-120.

Kraemer, M. U., Sinka, M. E., Duda, K. A., Mylne, A. Q., Shearer, F. M., Barker, C. M., . . . Hay, S. I. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *Elife, 4*, e08347. doi: 10.7554/eLife.08347

Kramer, L. D. (2016). Complexity of virus-vector interactions. *Curr Opin Virol, 21*, 81-86. doi: S1879-6257(16)30104-3 [pii]
10.1016/j.coviro.2016.08.008

Kramer, L. D., & Ciota, A. T. (2015). Dissecting vectorial capacity for mosquito-borne viruses. *Curr Opin Virol, 15*, 112-118. doi: 10.1016/j.coviro.2015.10.003
S1879-6257(15)00153-4 [pii]

Lan, Q., & Fallon, A. M. (1990). Small heat shock proteins distinguish between two mosquito species and confirm identity of their cell lines. *Am J Trop Med Hyg, 43*(6), 669-676.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi: 10.1186/gb-2009-10-3-r25
gb-2009-10-3-r25 [pii]

Lee, A., Nolan, A., Watson, J., & Tristem, M. (2013). Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philos Trans R Soc Lond B Biol Sci, 368*(1626), 20120503. doi: 10.1098/rstb.2012.0503
rstb.2012.0503 [pii]

Lesbats, P., Engelman, A. N., & Cherepanov, P. (2016). Retroviral DNA Integration. *Chem Rev, 116*(20), 12730-12757. doi: 10.1021/acs.chemrev.6b00125

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

Maringer, K., Yousuf, A., Heesom, K. J., Fan, J., Lee, D., Fernandez-Sesma, A., . . . Davidson, A. D. (2017). Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in Aedes aegypti. *BMC Genomics, 18*(1), 101. doi: 10.1186/s12864-016-3432-5
10.1186/s12864-016-3432-5 [pii]

Miesen, P., Girardi, E., & van Rij, R. P. (2015). Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. *Nucleic Acids Res, 43*(13), 6545-6556. doi: 10.1093/nar/gkv590
gkv590 [pii]

Miesen, P., Ivens, A., Buck, A. H., & van Rij, R. P. (2016). Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs. *PLoS Negl Trop Dis, 10*(2), e0004452. doi: 10.1371/journal.pntd.0004452

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PNTD-D-15-01748 [pii]

Miesen, P., Joosten, J., & van Rij, R. P. (2016). PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. *PLoS Pathog, 12*(12), e1006017. doi: 10.1371/journal.ppat.1006017
PPATHOGENS-D-16-02114 [pii]

Mongelli, V., & Saleh, M. C. (2016). Bugs Are Not to Be Silenced: Small RNA Pathways and Antiviral Responses in Insects. *Annu Rev Virol, 3*(1), 573-589. doi: 10.1146/annurev-virology-110615-042447

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. J., . . . Severson, D. W. (2007). Genome sequence of Aedes aegypti, a major arbovirus vector. *Science, 316*(5832), 1718-1723. doi: 1138878 [pii]
10.1126/science.1138878

Parrish, N. F., Fujino, K., Shiromoto, Y., Iwasaki, Y. W., Ha, H., Xing, J., . . . Tomonaga, K. (2015). piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. *RNA, 21*(10), 1691-1703. doi: 10.1261/rna.052092.115
rna.052092.115 [pii]

Peleg, J. (1968). Growth of arboviruses in monolayers from subcultured mosquito embryo cells. *Virology, 35*(4), 617-619.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol, 26*(7), 1641-1650. doi: 10.1093/molbev/msp077
msp077 [pii]

Rosenkranz, D., Rudloff, S., Bastuck, K., Ketting, R. F., & Zischler, H. (2015). Tupaia small RNAs provide insights into function and evolution of RNAi-based transposon defense in mammals. *RNA, 21*(5), 911-922. doi: 10.1261/rna.048603.114
rna.048603.114 [pii]

Rosenkranz, D., & Zischler, H. (2012). proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics, 13*, 5. doi: 10.1186/1471-2105-13-5
1471-2105-13-5 [pii]

Servant, G., & Deininger, P. L. (2015). Insertion of Retrotransposons at Chromosome Ends: Adaptive Response to Chromosome Maintenance. *Front Genet, 6*, 358. doi: 10.3389/fgene.2015.00358

Staginnus, C., Gregor, W., Mette, M. F., Teo, C. H., Borroto-Fernandez, E. G., Machado, M. L., . . . Schwarzacher, T. (2007). Endogenous pararetroviral sequences in tomato (Solanum lycopersicum) and related species. *BMC Plant Biol, 7*, 24. doi: 1471-2229-7-24 [pii]
10.1186/1471-2229-7-24

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics, 27*(7), 1009-1010. doi: 10.1093/bioinformatics/btr039
btr039 [pii]

Thompson, P. J., Macfarlan, T. S., & Lorincz, M. C. (2016). Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Mol Cell, 62*(5), 766-776. doi: 10.1016/j.molcel.2016.03.029
S1097-2765(16)30012-0 [pii]

Vicoso, B., & Bachtrog, D. (2015). Numerous transitions of sex chromosomes in Diptera. *PLoS Biol, 13*(4), e1002078. doi: 10.1371/journal.pbio.1002078
PBIOLOGY-D-14-02861 [pii]

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Watson, M., Schnettler, E., & Kohl, A. (2013). viRome: an R package for the visualization and analysis of viral small RNA sequence datasets. *Bioinformatics, 29*(15), 1902-1903. doi: 10.1093/bioinformatics/btt297
btt297 [pii]

Whitehorn, J., Kien, D. T., Nguyen, N. M., Nguyen, H. L., Kyrylos, P. P., Carrington, L. B., . . . Simmons, C. P. (2015). Comparative Susceptibility of Aedes albopictus and Aedes aegypti to Dengue Virus Infection After Feeding on Blood of Viremic Humans: Implications for Public Health. *J Infect Dis, 212*(8), 1182-1190. doi: 10.1093/infdis/jiv173
jiv173 [pii]

Yamanaka, S., Siomi, M. C., & Siomi, H. (2014). piRNA clusters and open chromatin structure. *Mob DNA, 5*, 22. doi: 10.1186/1759-8753-5-22
1759-8753-5-22 [pii]

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

LEGENDS FOR MAIN FIGURES

Figure 1. Contiguity of the *Aedes aegypti* genome is drastically improved in the Aag2 assembly. (A) Histogram of contig length vs. total amount of sequence contained in each bin. The Aag2 assembly achieved the largest contig sizes (by an order of magnitude) compared to previous *Aedes aegypti*-derived assemblies. This large contig size also resulted in more overall sequence information/number of bases. (B) Boxplots indicating number of contigs aligned between LVP (Sanger) and Aag2 (PacBio). When aligned to each other, more contigs from LVP are aligned to larger Aag2 contigs than vice versa. (C) Dot plot alignment of multiple contigs from LVP (each denoted by a different color) to a single contig from Aag2 (sequenced by Pac Bio). Expanded repeats in the Aag2 assembly can be seen where the same LVP contig aligns to multiple Aag2 locations. In some instances, portions of multiple LVP contigs align to a single locus within the Aag2 contig (vertical ‘spikes’ of alignment). (D) More detailed dot plot alignments between Aag2 and LVP assemblies. Regions in the LVP assembly are expanded within the Aag2 assembly at numerous loci. Regions corresponding to putative piClusters in the Aag2 assembly are highlighted with grey shaded boxes. Each panel is labeled with the contig and position of the predicted piCluster shown.

Figure 2. The repeat landscape of the *Aedes aegypti* genome is predominantly made up of transposable elements. (A) Circular plot of the Aag2 assembly, showing the 10 largest contigs (black rectangles; ordered by size) and transposable elements (circles; colored by TE class). Transposable elements are prevalent throughout the congtigs (and the entire *aegypti* genome). Rectangles representing contigs are staggered to indicate relative contig size. (B) Density plot showing the distribution of TE content for contigs in the assembly. Red line indicates the genome-wide TE content. (C) Density plot comparing TE counts binned by TEs per contig between the Aag2, LVP(Sanger), and UCB (Illumina) assemblies. The Aag2 assembly has significantly more TEs per single contig than previous assemblies. (D) Pie chart representing TE class representation in the Aag2 genome. Further detail can be found in Table 2. (E) Stacked area plot showing the relative density of TEs along an example contig. Local proportions are based on a window size of 25 kbps. (F) Stacked histogram of Kimura divergence for classes of TEs found in the Aag2 assembly, expressed as a function of percentage of the genome. A relatively recent expansion/active phase of LTRs is evident (increase in LTRs at low Kimura divergence scores). Kimura divergence scores are based on the accumulated mutations of a given TE sequence compared to a consensus.

Figure 3. Endogenous Viral Elements (EVEs) are found throughout the Aag2 genome and are observed in piClusters. (A) Circular plot showing piRNAs mapping to the Aag2 genome, with those derived from EVEs highlighted in red. Inner tracks show EVEs found in the Aag2 genome, colored by virus family from which they derived. (B) i) Density plot of TEs along Aag2 contig 000015F. Local enrichment for particular TE classes can be seen. ii) Zoom of

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

indicated region in i). Transposons are indicated by colored arrows, categorized by TE class. '+' and '-' indicate strandedness of the transposons. EVEs are shown as black triangles, organized by strandedness. piRNAs mapping to the region are indicated at the bottom, in either sense or anti-sense direction. Clear clusters of EVEs and TEs can be seen, with piRNAs specifically mapping to these clusters. The further zoom (iii) better shows the shared directionality of EVEs (colored triangles) and their surrounding transposon sequences (black, semi-transparent triangles). (C) Bar plot showing counts of EVEs derived from different viral families.(D) Bar plot showing a larger number of piRNA reads can be successfully mapped to the Aag2 (PacBio) assembly compared to the LVP (Sanger) assembly.

Figure 4. EVEs are primarily associated with LTR transposable elements. (A) Histograms showing counts of non-overlapping TEs closest to EVEs binned by distance, both upstream (negative x-axis values) and downstream (positive x-axis values). Positive y-axis counts refer to TE/EVE 'pairs' with the same strandedness , while negative y-axis counts are EVEs where the closest TE has the opposite strandedness. The "+/-" value indicates the ratio of TE/EVE pairs with the same strandedness to those with the opposite strandedness. Total counts represented in each histogram: All classes (n=766); LTR only (n=475); No LTRs (n=291); Ty3/gypsy only (n=274); Pao Bel only (n=180); Ty1/copia only (n=21). (B) Heatmap showing categories of TEs nearest EVEs, categorized by the viral family from which the EVEs were derived. Only TEs with the same strandedness as its nearest EVE are shown. A "*" indicates significant enrichment by one-sided binomial test against the background prevalence of a given TE category in the genome (eg among all LTRs nearest Chuviridae-derived EVEs, Pao Bel elements are specifically enriched compared to the genome-wide counts of Pao Bel among all LTRs). Color indicates proportion of a given TE category nearest EVEs derived from the indicated viral family. Grey indicates the element was not found to be the closest TE to any EVEs derived from the indicated viral family. Only TE elements which made up at least 10% of the dataset for a given viral family are shown. "Pao Bel elements" refers to Chuviridae, while "Ty3/gypsy elements" corresponds to Flaviviridae and Rhabdoviridae. Total sample size of all TEs analyzed for each dataset: AllTEs-Rhabdoviridae (n=223),Flaviviridae (n=118), Chuviridae (n=112); RNA-basedTEs- Rhabdoviridae (n=160),Flaviviridae (n=116), Chuviridae (n=95); LTRs- Rhabdoviridae (n=123),Flaviviridae (n=106), Chuviridae (n=84); Pao Bel-Chuviridae (n=67); Ty3/gypsy- Rhabdoviridae (n=99), Flaviviridae (n=77).

Figure 5. EVEs found in the genomes of *Aedes aegypti* and *Aedes albopictus* are derived from overlapping families of viruses, but do not share the same sequences. (A) Phylogenetic relationship between 61 members of Flaviviridae. EVEs present in (i) *Ae. aegytpi* or (ii) *Ae. albopictus* which align to the indicated virus are marked with a colored circle. Size corresponds to abundance of EVEs derived from given species. (B) Coverage plots of EVEs derived from the viral families (i) Flaviviridae, (ii) Rhabdoviridae, and (iii)

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Chuviridae. Each bar represents a single EVE, while its length and position denotes the region of the indicated ORF from which its sequence is derived. Length is expressed as a percentage of the total ORF, in order to normalize for varying ORF lengths among different members of a given viral family. In (i), the genome of CFAV is presented for reference. In (ii) and (iii), a generic genome is presented to better illustrate where EVEs are derived from within the genome as a whole (and within each specific ORF). (C) Violin plot of Kimura divergence scores for EVEs' nearest-neighbor TEs. Only EVEs whose nearest neighbor TE is non-overlapping and shares the same strandedness are shown. Both upstream and downstream nearest neighbors are represented. Counts in each category are as follows: Pao Bel-All (n=113,526); Pao Bel-Chuviridae (n=36); Pao Bel-Flaviviridae (n=12); Pao Bel-Rhabdoviridae (n=14); Ty3/gypsy -All (n=174,353); Ty3/gypsy -Chuviridae (n=10); Ty3/gypsy -Flaviviridae (n=43); Ty3/gypsy-Rhabdoviridae (n=53). Differences between indicated distributions were determined by Kolmogorov-Smirnov test. '*': p-value <0.05; '**': p-value <5e-4; '***': p-value <5e-8.

bioRxiv preprint doi: https://doi.org/10.1101/127498; this version posted April 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

LEGENDS FOR SUPPLEMENTARY FIGURES

Figure 2-Figure Supplement 1. Transposons are distributed throughout the entire Aag2 genome. (A) Similar plot to Figure 2A, but showing all contigs of the Aag2 assembly. Circular plot of the Aag2 assembly, with every contig (black rectangles; ordered by size) and transposable element (circles; colored by TE class). Transposable elements are prevalent throughout the entire *Ae. aegypti* genome. Rectangles representing contigs are staggered to indicate relative contig size.

Figure 4-Figure Supplement 1. TEs which overlap EVEs are also overrepresented by LTR elements. (A) Histograms of TEs which overlap EVEs, broken down by the indicated categories. The left bin represents TEs whose start is upstream, and end overlaps the EVE. The right bin indicates TEs whose end is downstream, and start overlaps an EVE. The middle bin indicates TEs whose coordinates surround an EVE. Positive count values indicate TE and EVEs with shared directionality, while negative values represent TE and EVEs with opposite directionality. Some EVEs showed multiple overlapping TEs, all of which are represented on the charts. (B) Heatmap, as in Figure 3, showing EVE-overlapping TE ‘preference’ for Rhabodoviridae, Flaviviridae, and Chuviridae-derived EVEs.

Figure 3-Figure Supplement 1. EVEs are typically found within unidirectional piRNA clusters. The left panels correspond to a region of Contig 000933F encoding 4 tandem, unidirectional piRNA clusters (as identified by proTRAC), each containing EVEs. Each cluster expresses piRNAs primarily anti-sense to the TEs/EVEs which define them. Similarly, a single large piRNA cluster on Contig 000044F is shown in the right panels. The shared directionality between TEs and EVEs (Figure 3B) is evident. Again, piRNA expression is almost exclusively in the antisense direction with respect to the TEs/EVEs.

Figure 2-Figure Supplement 2. Kimura divergence scores of LTRs only show expansion of Pao Bel and Ty3/gypsy elements. Bar plot of kimura scores assigned to LTRs only, categorized by TE family and expressed as percent of total genome (as in Figure 2E). At very low (0-1) Kimura divergence scores, Ty3/gypsy and Pao Bel exhibit a marked increase in proportion of the genome.

Figure 5-Figure Supplement 1. Kimura divergence of EVE proximal TEs is distinct from TEs genome wide. Density plots of kimura distributions of all Pao Bel or Ty3/gypsy TEs and EVE-proximal TEs. EVE-proximal TEs are further categorized by the viral family of the EVE it is nearest.

	UCB	LVP	Aag2
Sample	LVP strain	LVP strain	Aag2 cell line
Seq. Strategy	Illumina	Sanger	PacBio
Released	5/2015	6/2006	NA
Coverage	6.8x	7.6x	~50x
Total sequence length	744,596,036	1,383,957,531	1,723,930,323
Total assembly gap length	196,533,049	73,881,199	0
Num. of contigs	961,292	36,204	3,752
Contig N50	989	82,618	1,420,116
Contig L50	151,087	4,346	368

	Num. of elements	Length (Mbp)	Percent of genome
SINE	28,301	4.4	0.25
LINE	558,382	259.9	15.07
LTR	495,204	163.9	9.51
DNA	1,184,522	309.0	17.93
Other*	725,958	233.7	13.55
Total	2,992,367	970.9	56.31

*includes helitrons, MITEs, Penelope, RC, UD, and unknown elements

Fig. 1

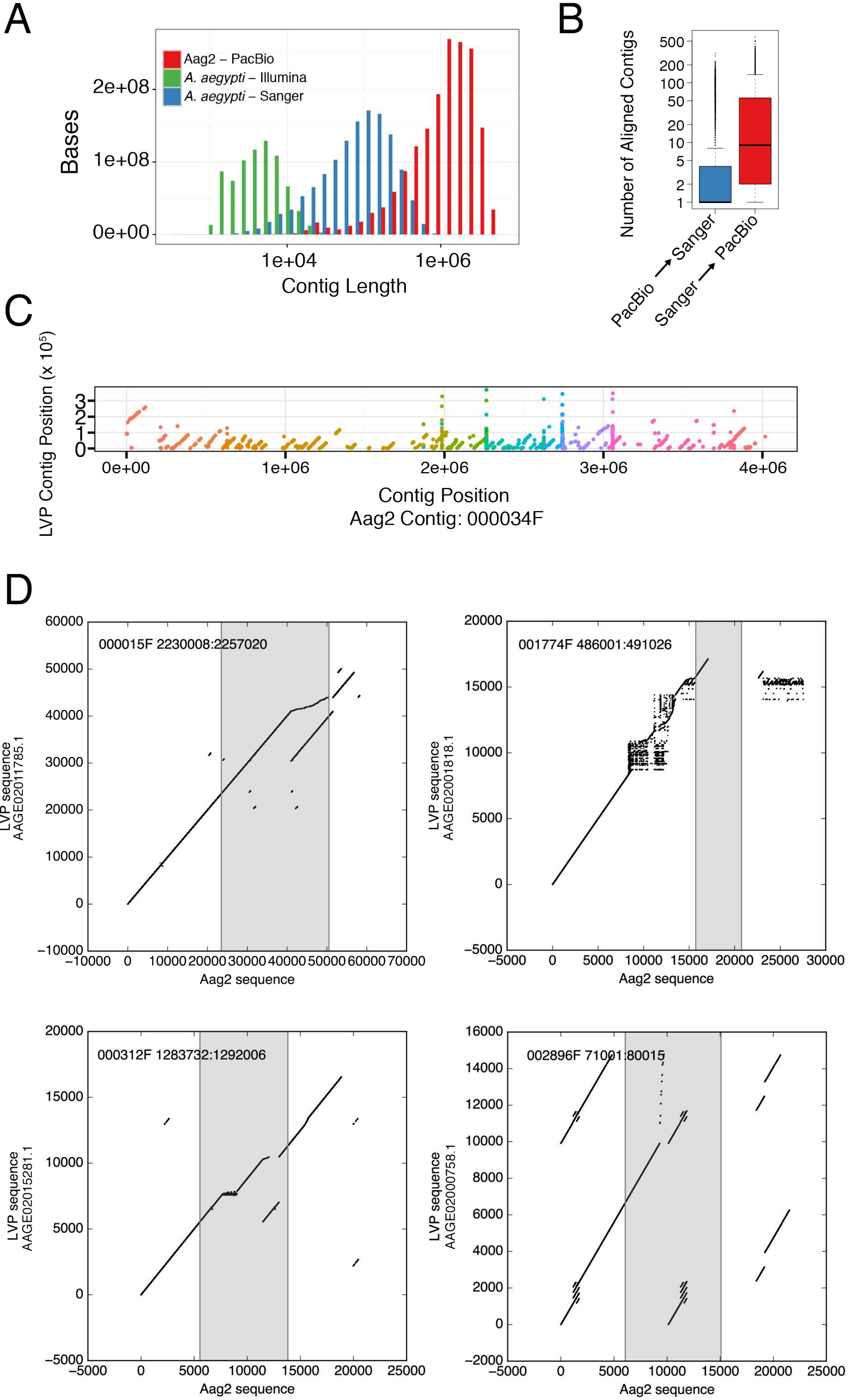

A
Aag2 – PacBio
A. aegypti – Illumina
A. aegypti – Sanger
Bases
2e+08
1e+08
0e+00
1e+04
1e+06
Contig Length
B
Number of Aligned Contigs
500
200
100
50
20
10
5
2
1
PacBio → Sanger
Sanger → PacBio
C
LVP Contig Position (x 10⁵)
3
2
1
0
0e+00
1e+06
2e+06
3e+06
4e+06
Contig Position
Aag2 Contig: 000034F
D
000015F 2230008:2257020
LVP sequence AAGE02011785.1
Aag2 sequence
001774F 486001:491026
LVP sequence AAGE02001818.1
Aag2 sequence
000312F 1283732:1292006
LVP sequence AAGE02015281.1
Aag2 sequence
002896F 71001:80015
LVP sequence AAGE02000758.1
Aag2 sequence

Fig. 2

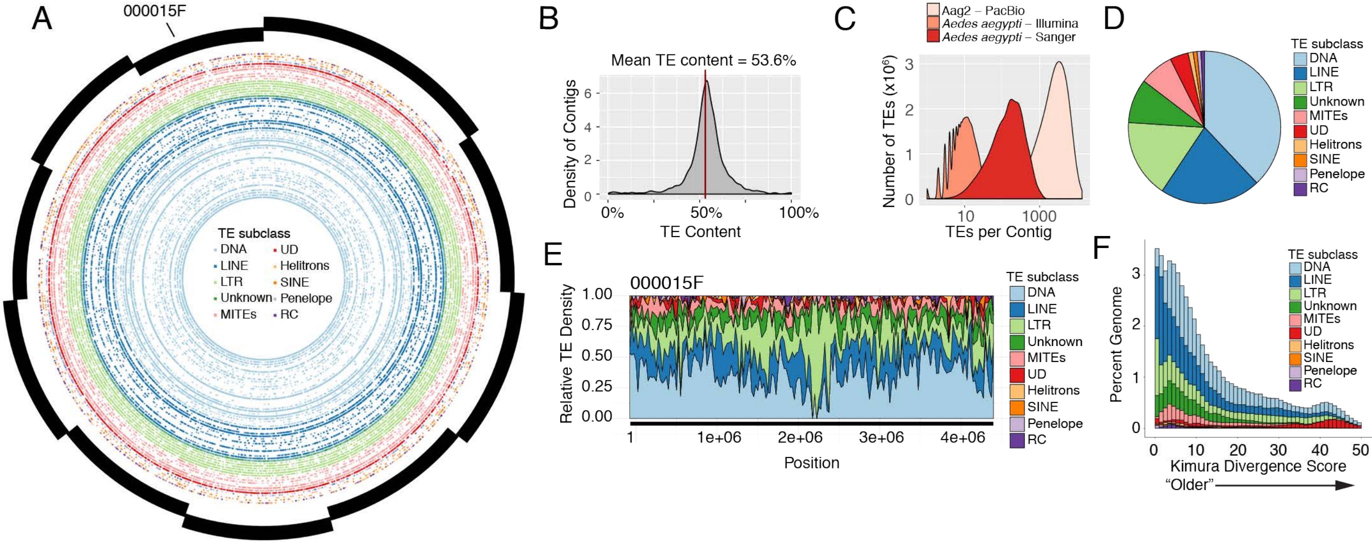

A
000015F
TE subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
B
Mean TE content = 53.6%
Density of Contigs
0
2
4
6
0%
50%
100%
TE Content
C
Aag2 – PacBio
Aedes aegypti – Illumina
Aedes aegypti – Sanger
Number of TEs (x10⁶)
0
1
2
3
10
1000
TEs per Contig
D
TE subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
E
000015F
Relative TE Density
0.00
0.25
0.50
0.75
1.00
1
1e+06
2e+06
3e+06
4e+06
Position
TE subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
F
Percent Genome
0
1
2
3
0
10
20
30
40
50
Kimura Divergence Score
"Older"
TE subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC

Fig. 3

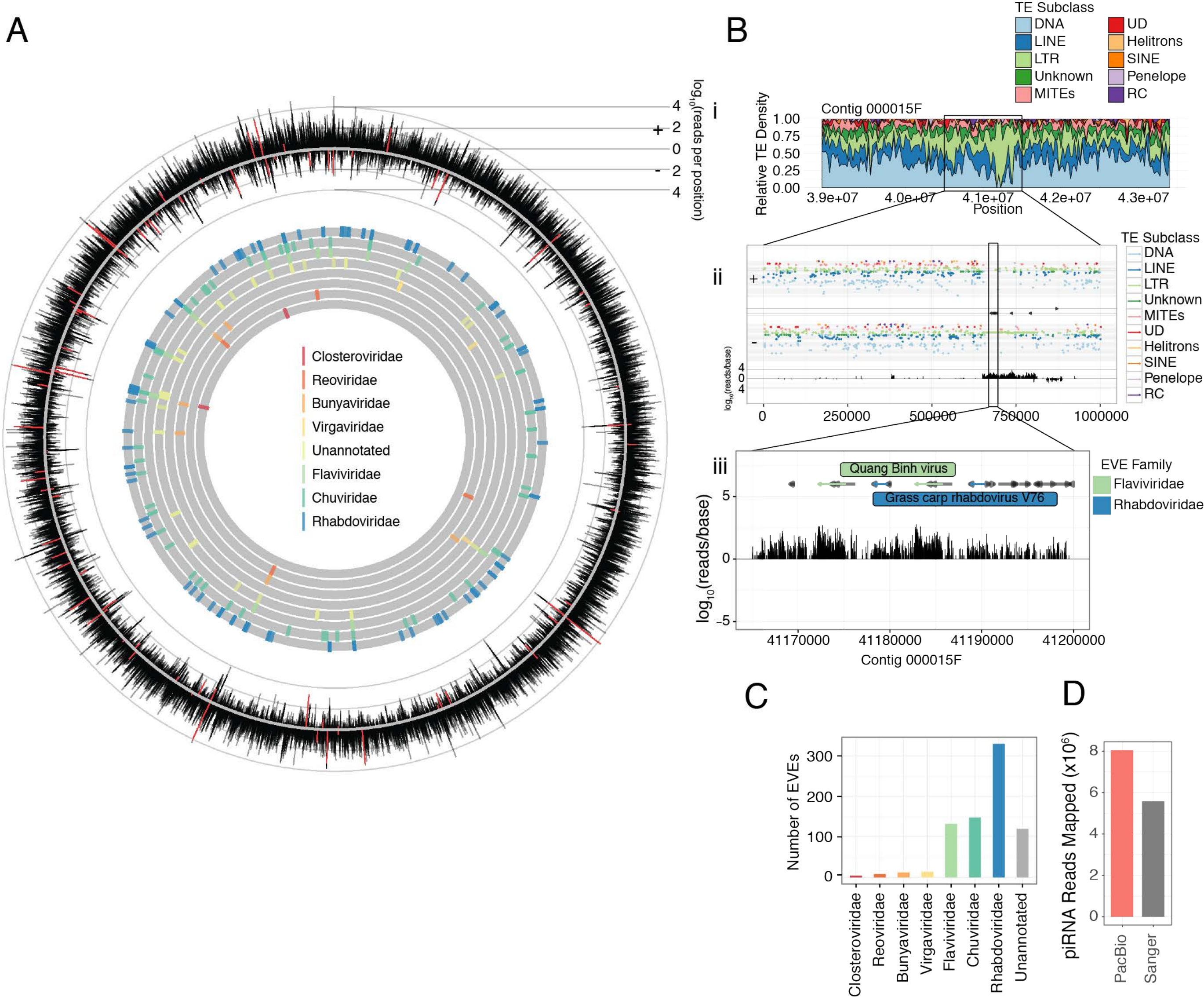

A
4
+ 2
0
- 2
4
log10(reads per position)
Closteroviridae
Reoviridae
Bunyaviridae
Virgaviridae
Unannotated
Flaviviridae
Chuviridae
Rhabdoviridae
B
TE Subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
i
Contig 000015F
Relative TE Density
1.00
0.75
0.50
0.25
0.00
3.9e+07
4.0e+07
4.1e+07
4.2e+07
4.3e+07
Position
ii
+
-
log10(reads/base)
4
0
4
0
250000
500000
750000
1000000
iii
Quang Binh virus
Grass carp rhabdovirus V76
EVE Family
Flaviviridae
Rhabdoviridae
log10(reads/base)
5
0
-5
41170000
41180000
41190000
41200000
Contig 000015F
C
Number of EVEs
300
200
100
0
Closteroviridae
Reoviridae
Bunyaviridae
Virgaviridae
Flaviviridae
Chuviridae
Rhabdoviridae
Unannotated
D
piRNA Reads Mapped (x10^6)
8
6
4
2
0
PacBio
Sanger

Fig. 4

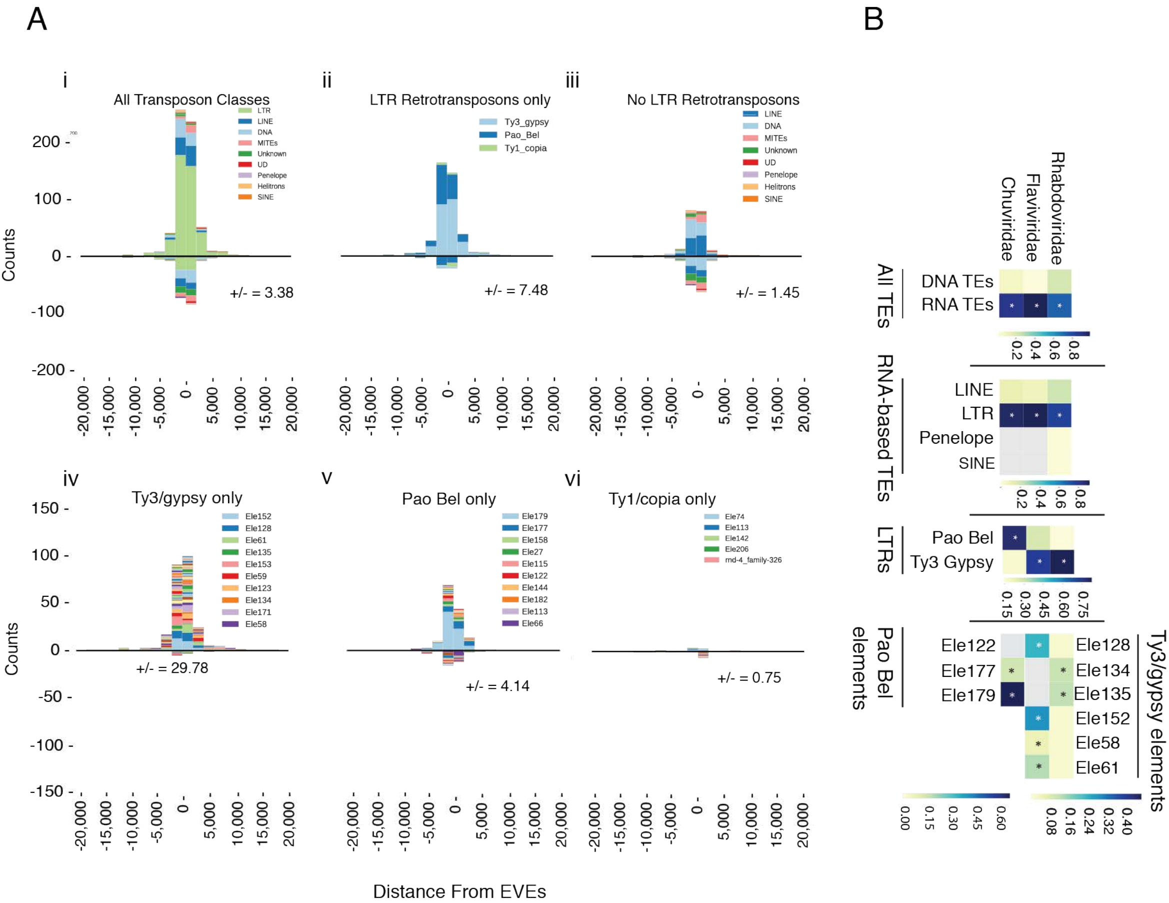

A
B
i
All Transposon Classes
LTR
LINE
DNA
MITEs
Unknown
UD
Penelope
Helitrons
SINE
+/- = 3.38
ii
LTR Retrotransposons only
Ty3_gypsy
Pao_Bel
Ty1_copia
+/- = 7.48
iii
No LTR Retrotransposons
LINE
DNA
MITEs
Unknown
UD
Penelope
Helitrons
SINE
+/- = 1.45
Counts
200
100
0
-100
-200
-20,000
-15,000
-10,000
-5,000
0
5,000
10,000
15,000
20,000
iv
Ty3/gypsy only
Ele152
Ele128
Ele61
Ele135
Ele153
Ele59
Ele123
Ele134
Ele171
Ele58
+/- = 29.78
v
Pao Bel only
Ele179
Ele177
Ele158
Ele27
Ele115
Ele122
Ele144
Ele182
Ele113
Ele66
+/- = 4.14
vi
Ty1/copia only
Ele74
Ele113
Ele142
Ele206
rnd-4_family-326
+/- = 0.75
Counts
150
100
50
0
-50
-100
-150
Distance From EVEs
Chuviridae
Flaviviridae
Rhabdoviridae
All TEs
DNA TEs
RNA TEs
0.2
0.4
0.6
0.8
RNA-based TEs
LINE
LTR
Penelope
SINE
0.2
0.4
0.6
0.8
LTRs
Pao Bel
Ty3 Gypsy
0.15
0.30
0.45
0.60
0.75
Pao Bel elements
Ele122
Ele177
Ele179
Ty3/gypsy elements
Ele128
Ele134
Ele135
Ele152
Ele58
Ele61
0.00
0.15
0.30
0.45
0.60
0.08
0.16
0.24
0.32
0.40
*

Fig. 5

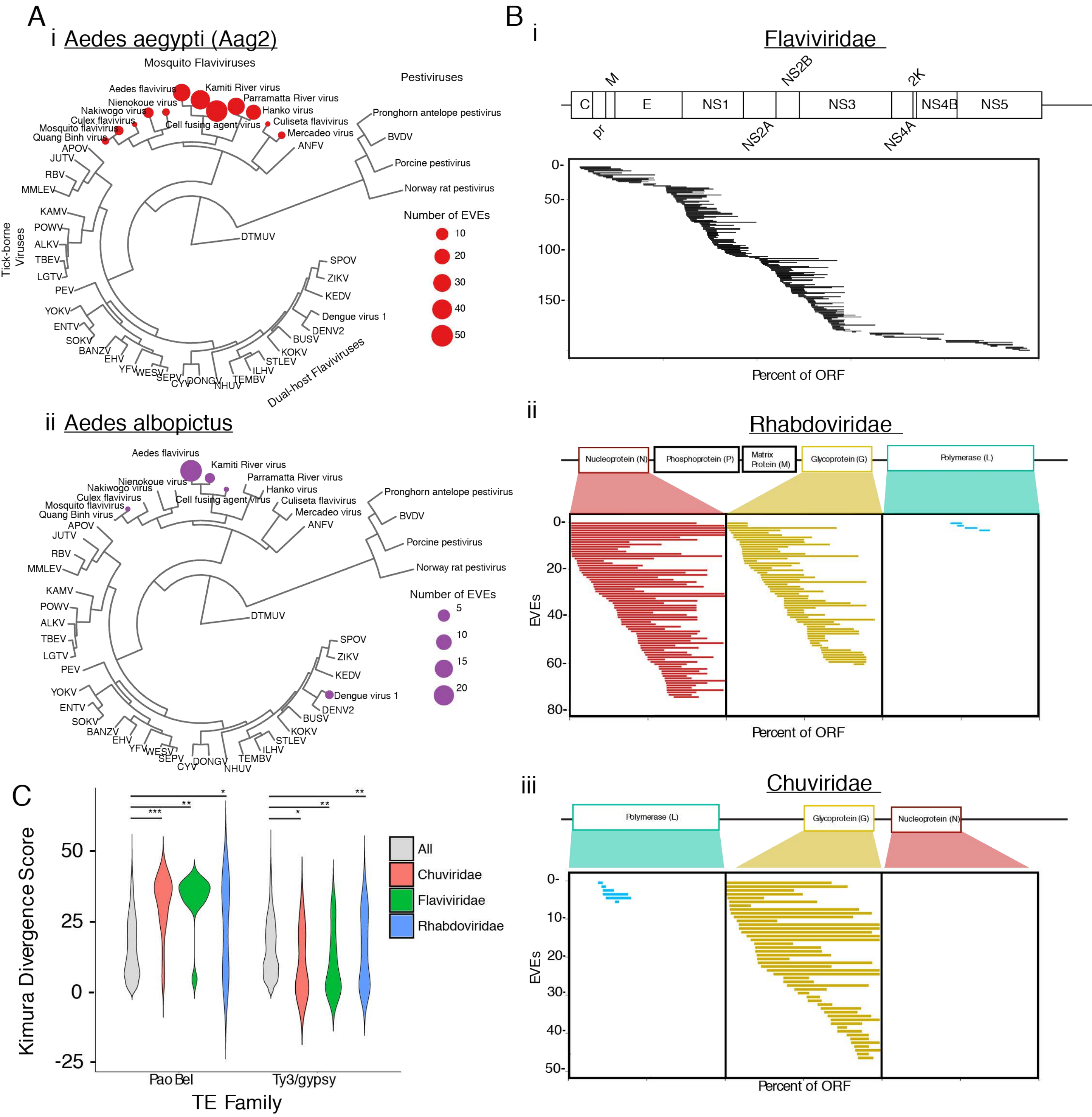

A
i Aedes aegypti (Aag2)
Mosquito Flaviviruses
Pestiviruses
Aedes flavivirus
Kamiti River virus
Nienokoue virus
Parramatta River virus
Nakiwogo virus
Hanko virus
Culex flavivirus
Cell fusing agent virus
Culiseta flavivirus
Mosquito flavivirus
Mercadeo virus
Quang Binh virus
ANFV
APOV
JUTV
RBV
MMLEV
KAMV
POWV
ALKV
TBEV
LGTV
PEV
Tick-borne Viruses
YOKV
ENTV
SOKV
BANZV
EHV
YFV
WESV
SEPV
CYV
DONGV
NHUV
TEMBV
ILHV
STLEV
KOKV
BUSV
DENV2
Dengue virus 1
KEDV
ZIKV
SPOV
Dual-host Flaviviruses
DTMUV
Pronghorn antelope pestivirus
BVDV
Porcine pestivirus
Norway rat pestivirus
Number of EVEs
10
20
30
40
50
ii Aedes albopictus
Number of EVEs
5
10
15
20
C
Kimura Divergence Score
50
25
0
-25
***
**
*
*
**
**
Pao Bel
Ty3/gypsy
TE Family
All
Chuviridae
Flaviviridae
Rhabdoviridae
B
i
Flaviviridae
C
pr
M
E
NS1
NS2A
NS2B
NS3
NS4A
2K
NS4B
NS5
0-
50-
100-
150-
Percent of ORF
ii
Rhabdoviridae
Nucleoprotein (N)
Phosphoprotein (P)
Matrix Protein (M)
Glycoprotein (G)
Polymerase (L)
EVEs
0-
20-
40-
60-
80-
Percent of ORF
iii
Chuviridae
Polymerase (L)
Glycoprotein (G)
Nucleoprotein (N)
EVEs
0-
10-
20-
30-
40-
50-
Percent of ORF

Fig. 2, Suppl 1

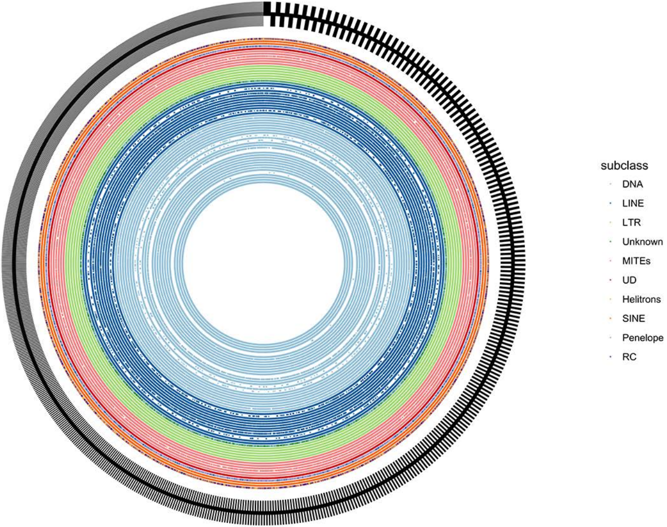
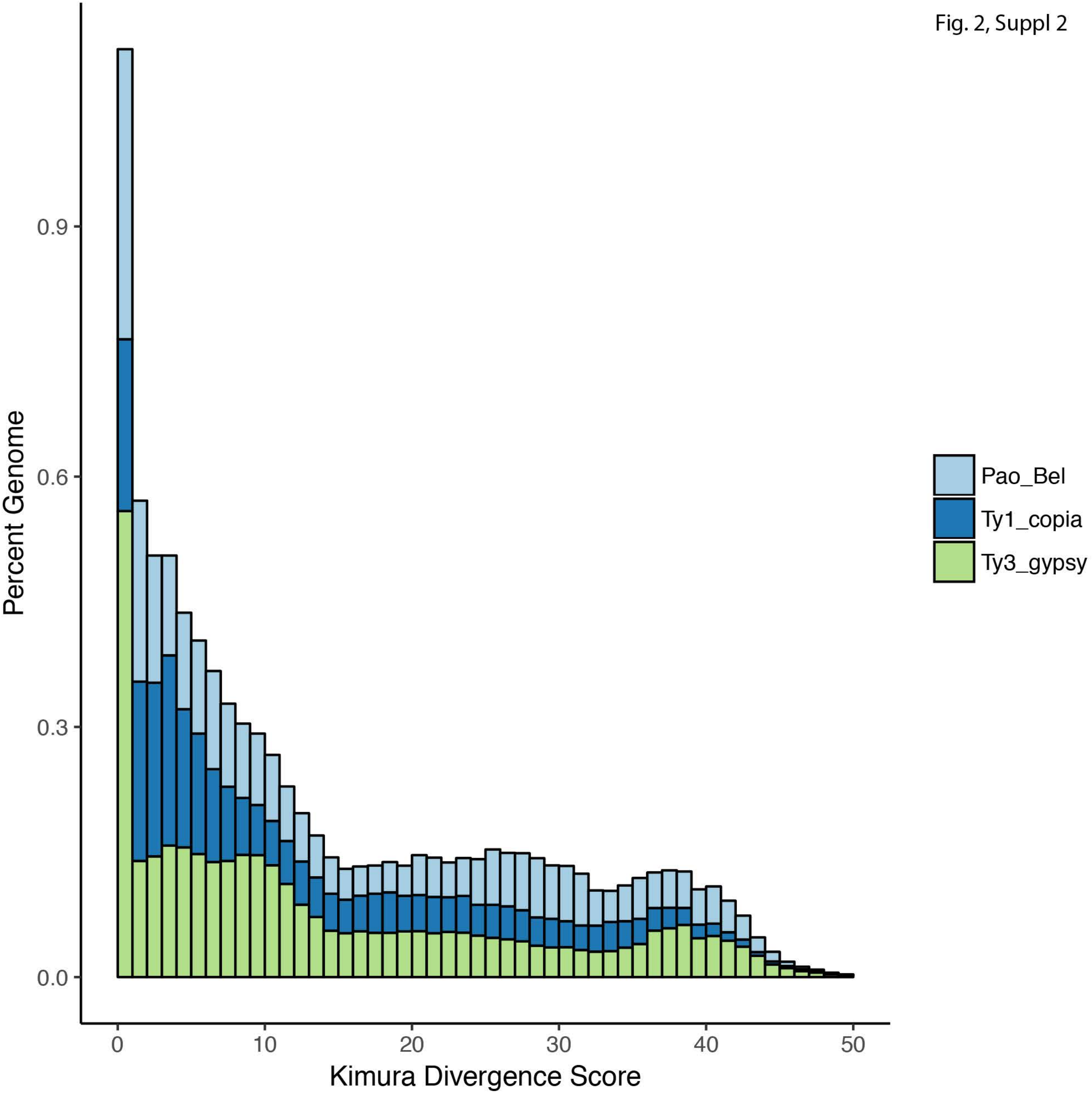
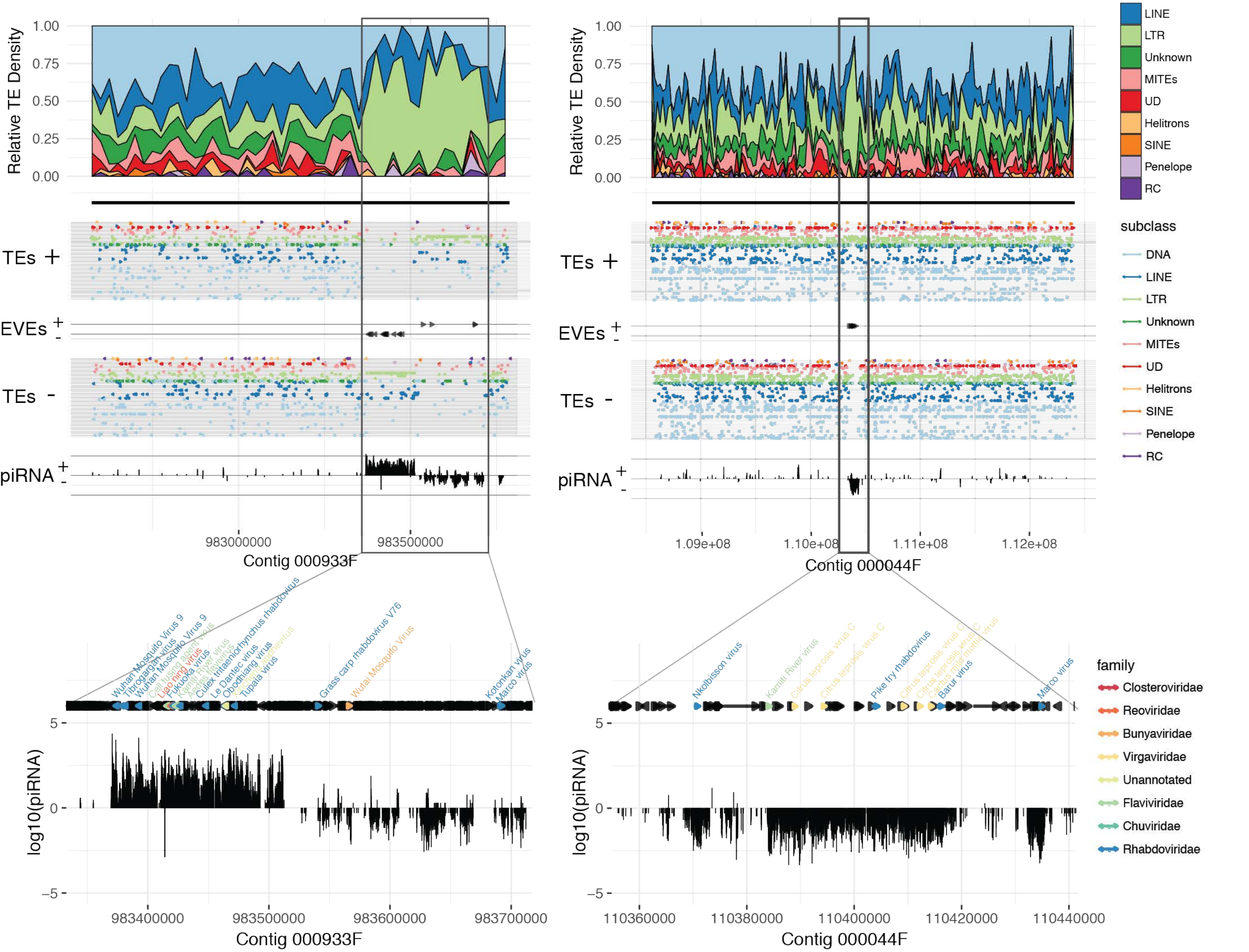

subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC

Fig. 2, Suppl 2



Percent Genome
0.0
0.3
0.6
0.9
0
10
20
30
40
50
Kimura Divergence Score
Pao_Bel
Ty1_copia
Ty3_gypsy

Fig. 3, Suppl 1



Relative TE Density
1.00
0.75
0.50
0.25
0.00
TEs +
EVEs +
-
TEs -
piRNA +
-
983000000
983500000
Contig 000933F
Relative TE Density
1.00
0.75
0.50
0.25
0.00
TEs +
EVEs +
-
TEs -
piRNA +
-
1.09e+08
1.10e+08
1.11e+08
1.12e+08
Contig 000044F
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
subclass
DNA
LINE
LTR
Unknown
MITEs
UD
Helitrons
SINE
Penelope
RC
Wuhan Mosquito Virus 9
Tibrogargan virus
Wuhan Mosquito Virus 9
Cell fusing agent virus
Liao ning virus
Fukuoka virus
Kamiti River virus
Aedes flavivirus
Culex tritaeniorhynchus rhabdovirus
Le Dantec virus
Obodhiang virus
Xincheng anphevirus
Tupaia virus
Grass carp rhabdovirus V76
Wutai Mosquito Virus
Kotonkan virus
Marco virus
5
0
-5
log10(piRNA)
983400000
983500000
983600000
983700000
Contig 000933F
Nkolbisson virus
Kamiti River virus
Citrus leprosis virus C
Citrus leprosis virus C
Pike fry rhabdovirus
Citrus leprosis virus C
Citrus leprosis virus C
Cactus mild mottle virus
Barur virus
Marco virus
5
0
-5
log10(piRNA)
110360000
110380000
110400000
110420000
110440000
Contig 000044F
family
Closteroviridae
Reoviridae
Bunyaviridae
Virgaviridae
Unannotated
Flaviviridae
Chuviridae
Rhabdoviridae

Fig. 4, Suppl 1

A

Overlapping TE/EVE pairs

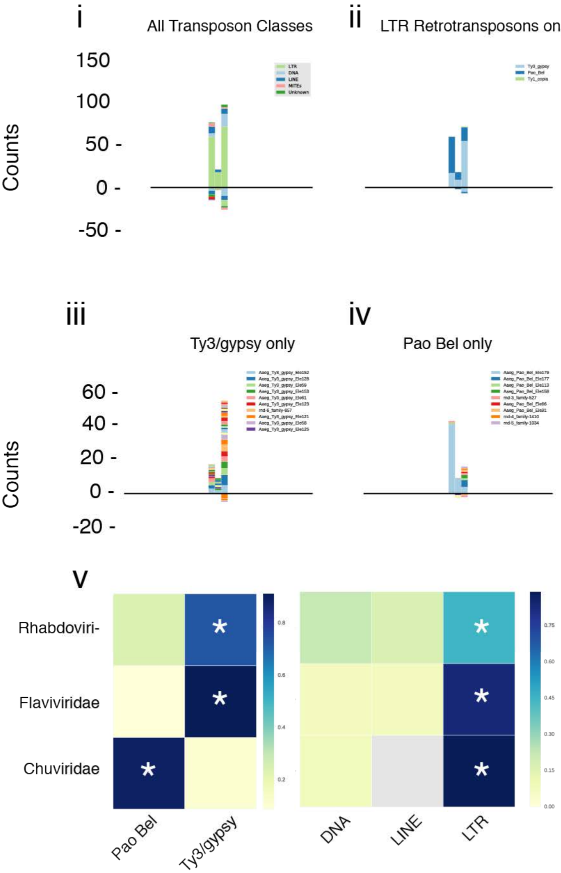
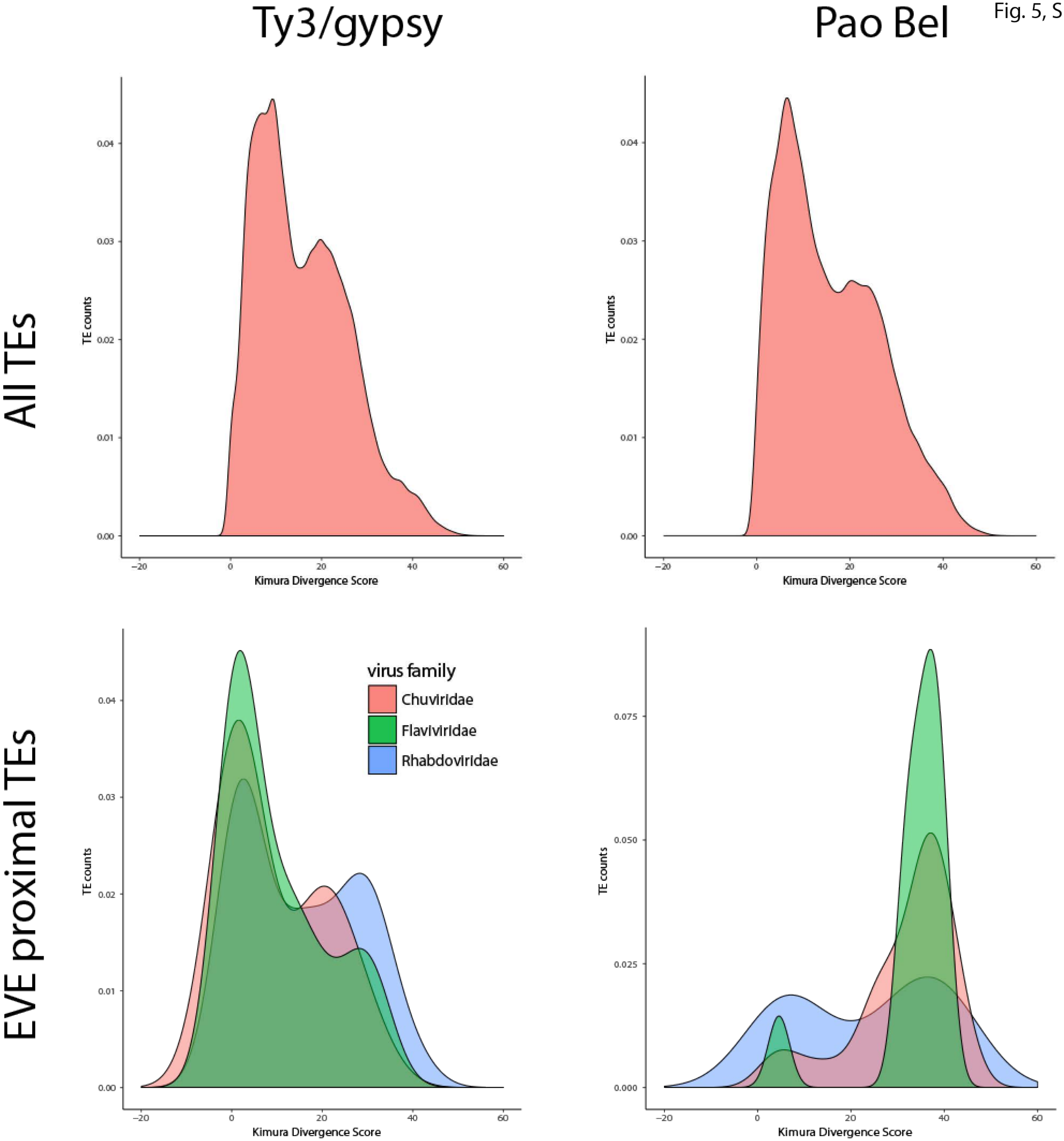


Fig. 5, Suppl 1


Ty3/gypsy
Pao Bel
All TEs
EVE proximal TEs
TE counts
0.00
0.01
0.02
0.03
0.04
−20
0
20
40
60
Kimura Divergence Score
TE counts
0.00
0.01
0.02
0.03
0.04
−20
0
20
40
60
Kimura Divergence Score
TE counts
0.00
0.01
0.02
0.03
0.04
−20
0
20
40
60
Kimura Divergence Score
virus family
🟥 Chuviridae
🟩 Flaviviridae
🟦 Rhabdoviridae
TE counts
0.000
0.025
0.050
0.075
−20
0
20
40
60
Kimura Divergence Score