Published in final edited form as:

Insect Mol Biol. 2015 June ; 24(3): 331–337. doi:10.1111/imb.12160.

Horizontal Transmission of an *R4* Clade Non-Long Terminal Repeat Retrotransposon between the Divergent *Aedes* and *Anopheles* Mosquito Genera

James K. Biedler¹, Xiaoguang Chen², and Zhijian Tu¹

¹Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

²Department of Pathogen Biology, School of Public Health and Tropical Medicine, Southern Medical University, Guang Zhou, Guang Dong 510515 P.R. China

Abstract

AaegR4_1 and *AgamR4_1* are the sole *R4* clade non-long terminal repeat retrotransposons in *Aedes aegypti* and *Anopheles gambiae*, two species that diverged approximately 145–200 million years ago. Twelve full-length copies were found in *Ae. aegypti* and have less than 1% nt divergence, suggesting recent activity on an evolutionary time scale. Five of these copies have intact open reading frames and the 3.6 kb open reading frame of *AaegR4_1.1* has 78% nt identity to *AgamR4_1.1*. No intact copies can be found in *An. gambiae*. Searches of 25 genomic databases for 22 mosquito species from 3 genera revealed *R4* clade representatives in *Aedes* and *Anopheles* genera but not in *Culex*. Interestingly, these elements are present in all 6 species of the *An. gambiae* species complex that were searched but not in 13 other Anopheline species. These results combined with divergence vs. age analysis suggest that horizontal transfer is the most likely explanation for the low divergence between *R4* clade retrotransposon sequences of the divergent mosquito species from the *Aedes* and *Anopheles* genera. This is the first report of the horizontal transfer of an *R4* clade non-LTR retrotransposon and the first report of the horizontal transfer of a non-LTR retrotransposon in mosquitoes.

Keywords

Horizontal Transfer; Mosquito; Evolution; Transposable Elements; Non-Long Terminal Repeat Retrotransposon

Please send correspondence to: Dr. James K. Biedler, Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, Telephone: 540-231-7395, Fax: 540-231-9070, jbiedler@vt.edu.
Co-author contact information: Dr. Zhijian Tu, Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, Telephone: 540-231-7395, Fax: 540-231-9070, jaketu@vt.edu
Xiao-Guang Chen (M.D & Ph.D), Head and Professor, Department of Pathogen Biology, Deputy Dean, School of Public Health and Tropical Medicine, Southern Medical University, Guang Zhou, Guang Dong 510515, P.R.China, xgchen2001@hotmail.com

Supporting Information

- S1. Information for *R4* clade sequences used in this study.
- S2. Sequences used for *white* gene phylogeny in Figure 2.
- S3. Table with supporting information for Figure 3.
- S4. Table with genome sequence databases searched showing host distribution of *R4* clade sequences.

Introduction

Transposable elements (TEs) are “selfish” genetic elements that are able to replicate and occupy significant portions of eukaryotic genomes (Doolittle and Sapienza, 1980). They are divided into Class I and Class II elements based on their mechanism of transposition (Finnegan, 1992). Non-long terminal repeat (non-LTR) retrotransposons are Class I TEs that replicate via an RNA intermediate utilizing a reverse transcriptase (RT). They form a clade that comprises divergent families differing in structural domains but share in common at least a reverse transcriptase with the exception of nonautonomous non-LTR retrotransposons. “Non-long terminal repeat” distinguishes this clade from long terminal repeat-containing retrotransposons that are distantly related. Non-LTR retrotransposons have been maintained in diverse taxonomic lineages (Malik et al., 1999) and extensive studies with the site-specific non-LTR retrotransposons *R1* and *R2* have shown that they have been vertically transmitted for long periods of evolutionary time (Eickbush and Eickbush, 1995, Lathe et al., 1995, Lathe and Eickbush, 1997, Kojima and Fujiwara, 2005). It has been proposed that non-LTR retrotransposons may not depend on horizontal transfer for their existence, unlike some Class II TEs (Malik et al., 1999, Eickbush and Malik, 2002). Malik, Burke, and Eickbush (1999) generated a plot of orthologous non-LTR retrotransposons RT divergence vs. age of host divergence that supported the vertical transmission of several diverse non-LTR retrotransposons. They included in their analysis three previously proposed cases of non-LTR retrotransposon horizontal transfer. Challenging those conclusions, they showed that values for these cases fit a regression line derived from vertically transmitted orthologous non-LTR retrotransposons (Malik et al., 1999, Eickbush and Malik, 2002). However, there are several reported cases of horizontal transfer of non-LTR retrotransposons with variable degrees of evidence (Zupunski et al., 2001, Kordis and Gubensek, 1999, Kordis and Gubensek, 1998, Novikova et al., 2007, Novikova et al., 2009, Sanchez-Gracia et al., 2005, Volff et al., 2000). In one review on transposable element horizontal transfer, a total of 14 non-LTR retrotransposon horizontal transfer cases were counted (Schaack et al., 2010).

R4 clade non-LTR retrotransposons have been identified in the silkworm *Bombyx mori* (Xiong and Eickbush, 1993), a nematode *Ascaris lumbricoides* (Burke et al., 1995), and teleost fish (Volff et al., 2001). Previously described *R4* clade elements have a single ORF encoding an RT and a restriction enzyme-like endonuclease that confers target site specificity. *B. mori* *Dong* inserts into nontranscribed ribosomal TAA repeats and *A. lumbricoides* *R4* inserts into 26S ribosomal RNA genes. *R4* clade elements have also been reported in the mosquito species *Ae. aegypti* and *An. gambiae* (Biedler and Tu, 2003, Kojima and Fujiwara, 2004). Here we report the horizontal transfer of an *R4* clade element, occurring between the lineages of *Ae. aegypti* and the *An. gambiae* species complex, lineages that diverged approximately 145 to 200 MYA (Krzywinski et al., 2006). We provide evidence of horizontal transfer based on nucleotide (nt) sequence identity, host species distribution, and TE divergence vs. host age analysis using RT domains, a method previously used to investigate the transmission of non-LTR retrotransposons (Malik et al., 1999, Zupunski et al., 2001, Kojima and Fujiwara, 2005, Eickbush and Malik, 2002, Sormacheva et al., 2012, Novikova et al., 2007). To our knowledge, this is the first report of

an *R4* clade non-LTR retrotransposon involved in horizontal transfer and the first report of the horizontal transfer of a non-LTR retrotransposon in mosquitoes.

Results and Discussion

AaegR4_1 and AgamR4_1, the *R4* clade families in *Ae. aegypti* and *An. gambiae*

AaegR4_1 and *AgamR4_1* are the only *R4* clade non-LTR retrotransposon families in *Ae. aegypti* and *An. gambiae*, respectively. No divergent *R4* clade-related sequences were found by BLASTn or TBLASTn (Altschul et al., 1990) using the *An. gambiae* *AgamR4_1* conceptually translated aa sequence as the query (Biedler and Tu, 2003). Search of the *Culex quinquefasciatus* (*C. quinquefasciatus*) genome assembly (Arensburger et al., 2010) did not reveal any *R4* clade representatives.

AaegR4_1 is the *Ae. aegypti* *R4* clade family that was found by a TBLASTn search of the *Ae. aegypti* genome assembly (Nene et al., 2007). Fourteen full-length *AaegR4_1* copies of 3994 bp (two are 3993 bp) were identified by aligning multiple copies and their flanking sequence with CLUSTAL (Thompson et al., 1997, Thompson et al., 1994). The full-length copy *AaegR4_1.1* is used in this study and its accession and coordinates are listed in S1. Twelve copies have greater than 99% nt identity to this sequence used as the query in BLAST vs. the *Ae. aegypti* genome AaegL3 assembly scaffolds (vectorbase.org). This sequence and an *An. gambiae* *R4* representative have been submitted to Tefam, a relational database designed for mosquito transposable elements (<http://tefam.biochem.vt.edu/tefam/index.php>). Many other BLAST hits of various lengths and high nt identity were also detected.

Full-length *AaegR4_1* copies are approximately 4 kb with a single 3.6 kb ORF encoding an RT and a restriction enzyme-like endonuclease (Figure 1). A comparison with other *R4* clade elements *Dong*, *R4*, and *Rex6*, (Volf et al., 2001, Xiong and Eickbush, 1993, Burke et al., 1995) confirm *AaegR4_1.1* as having a similar structure (not shown). The canonical AATAAA polyadenylation signal was not identified in any *AaegR4_1* copies. Full-length copies are flanked by TAA repeats, the same target site reported for *Dong* of *B. mori*. An alignment of the 12 full-length copies shows the 5' and 3' flanking sequences consisting of approximately 20 and 10 TAA repeats, respectively (not shown). No association with ribosomal RNA genes was found, contrary to what is known for *Ascaris lumbricoides* *R4* and *Bombyx mori* *Dong*. Approximately 65 copies can be detected in the genome sequence by BLASTn, which have at least 90% nt identity and 80% length compared to the 850 base pair (bp) query corresponding to the RT domain of *AaegR4_1.1* (Figure 1). More copies are not detected using a lower identity cutoff. Fifty-eight of the 65 copies are detected using a 95% identity cutoff. *AaegR4_1* appears to have been recently active in evolutionary time as the 12 full-length copies have pairwise nt distances of 1% or less. Five of these copies have intact open reading frames (ORFs) indicating they may be potentially active.

AgamR4_1, the *R4* clade family in the *An. gambiae* genome sequence (Holt et al., 2002), comprise more divergent and fragmented copies for which no intact sequences could be detected. There are 6 copies with 88–97% nt identity to the RT-encoding nt query. Like *AaegR4_1*, *AgamR4_1* copies appear to also have TAA repeats flanking their termini.

Because a full-length sequence could not be found in the genome assembly, we have used for our analyses the best available copy (*AgamR4_1.2*) as well as a previously reported reconstructed (not consensus or ancestral) sequence that has a complete ORF (Kojima and Fujiwara, 2004). We call this reconstructed copy *AgamR4_1.1*. Additional information for the above two *AgamR4_1* sequences, which are approximately 94% identical at the nt level, is given in S1. Comparative analyses between *AaegR4_1.1* and *AgamR4_1* were performed using both *AgamR4_1* sequences, which showed similar results. Therefore, for simplicity only analyses using *AgamR4_1.1* are described below except for 5' untranslated region (UTR) comparison (see below). Attempts to amplify a full-length sequence from *An. gambiae* genomic DNA (PEST strain) by PCR were unsuccessful, which is consistent with the lack of full-length copies in the genome assembly. Similarly, a search of the genome sequences for *An. gambiae* M, now called *An. coluzzii* (Coetzee et al., 2013), and S forms (vectorbase.org) found they were devoid of intact *AgamR4_1* coding sequences.

Distribution of mosquito R4 clade representatives in mosquitoes strongly supports horizontal transfer

One line of evidence for horizontal transfer is when the TE distribution in its host species is not consistent with host phylogeny. To demonstrate this we constructed a basic mosquito phylogeny including 7 species from 3 genera based on the *white* gene (Figure 2). Using BLASTn with *AaegR4_1.1* and *AgamR4_1.1* as queries to search 25 genomes for 22 species (Marinotti et al., 2013, Zhou et al., 2014) (see S4), *R4* clade representatives were found only in species from two of the three genera, *Aedes* and *Anopheles*. More interesting is that *AgamR4_1*-like representatives were only found in the *An. gambiae* species complex and not in any other *Anopheline* genomic databases searched (see S4). Six species in the *An. gambiae* species complex were positive but 13 species outside the complex were negative. Notably, *An. christyi* and *An. epiroticus*, which are very closely related to the *An. gambiae* species complex were negative. We also searched the raw reads of the *An. stephensi* and *C. quinquefasciatus* genomes by BLASTn and did not find any hits for either *Ae. aegypti* or *An. gambiae* *R4* clade representatives. We searched *An. christyi*, *An. epiroticus*, and *An. stephensi* genomes by BLAST using the flanking sequences of *AgamR4_1* copies and were able to find corresponding genomic sequences but they did not have *R4*-related sequence insertions nearby, supporting the absence of *R4*-related sequences in these genomes. Therefore, there is no evidence that failure to detect copies in these species is a result of them not being included in the assemblies. By using BLAST we also could not find any evidence that any *AgamR4* and *AaegR4* sequences are derived from orthologous insertions. These findings strongly support horizontal transfer between the aedine and anopheline lineages. More specifically, the transfer seems to have occurred between the lineages of the common ancestors of *Ae. aegypti*/*Ae. albopictus* and the *An. gambiae* species complex because the *Ae. aegypti* and *Ae. albopictus* sequences are more similar to each other than they are to the *Anopheles* sequences. To identify orthologous insertions we searched the *Ae. albopictus* genome sequence (unpublished, see S4) by BLAST using 173 *Ae. aegypti* sequences as queries that comprised approximately 100 nt of the 3' end of *Ae. aegypti* *R4*-related sequences and 100 nt of 3' genomic flanking sequences. However, while we were able to detect orthologous genomic sequences, we were unable to find orthologous insertions between *Ae. aegypti* and *Ae. albopictus*.

AaegR4_1 and AgamR4_1 have less divergence than expected for vertically transmitted non-LTR retrotransposons

The mosquito *R4* clade families in *Ae. aegypti* and *An. gambiae* exhibit unusually low divergence for sequences expected to be vertically inherited in their host genomes. The 3.6 Kb ORFs of *AaegR4_1.1* and *AgamR4_1.1* have 78% and 75% sequence identity at the nt and aa level. *AaegR4_1.1* and *AgamR4_1.2* have 208 bp and 162 bp 5' UTRs with 81% identity. *AgamR4_1.2* was used for the 5' UTR comparison because *AgamR4_1.1* does not have the complete 5' UTR. *AaegR4_1.1* and *AgamR4_1.1* have 138bp and 136bp 3'UTRs with 70% identity.

By plotting RT amino acid (aa) divergence of orthologous non-LTRs vs. their host's age (divergence time of their hosts), a regression line can be derived that can be a guide to pointing out candidates of horizontal transfer (Malik et al., 1999, Zupunski et al., 2001). To analyze mosquito non-LTR retrotransposons, 13 conceptually translated RT aa sequences representing all 10 non-LTR retrotransposon clades from *An. gambiae* non-LTR elements, most which had intact full-length copies (Biedler and Tu, 2003), were used for TBLASTn vs. the *Ae. aegypti* genome sequence. The best match was then used for divergence vs. age analysis (Figure 3). We realize that not all of these best-match pairs between *Ae. aegypti* and *An. gambiae* are necessarily orthologous. It is possible that the sequences compared are paralogs. Regardless, a non-LTR retrotransposon pair that falls significantly below the reference regression line is a candidate for horizontal transfer. In contrast, vertically transmitted paralogous non-LTR retrotransposons are expected to be above the reference regression line.

A plot of RT aa divergence vs. age of host divergence shows mosquito *R4* sequences far below the regression line, suggesting a horizontal transfer (Figure 3). A studentized deleted residual (RStudent) was used to estimate the significance of deviation from the regression line. RStudent is a residual of the i^{th} case calculated from a regression model that has excluded the i^{th} case, which is then divided by an estimate of its standard deviation. The RStudent value for mosquito *R4* clade sequences (4.0) is greater than 2, suggesting it may be an outlier, consistent with expectations for horizontal transfer. RStudent was also calculated using only the mosquito non-LTR retrotransposons in question, thus eliminating the influence from all other mosquito non-LTR retrotransposon comparisons that may alter the regression line due to paralogous comparisons or other cases of horizontally transferred or unusually conserved sequences. In this case RStudent increases from 4.0 to 5.7, even more strongly suggesting horizontal transfer. Interestingly, mosquito *RTE* clade elements have an RStudent of 2.3 and non-mosquito *RTE* elements have been implicated in horizontal transfer (see above). The *Outcast* comparison (RStudent = 1.6) may represent a paralogous comparison resulting in higher than expected divergence.

It may be that *AgamR4_1.1* (*DongAg*), being a reconstructed sequence, could bias the aa identity comparisons. Therefore we obtained the longest intact sequence from the *An. gambiae* genome corresponding to the *AgamR4_1* RT region. This copy is called *AgamR4_1.2* (see Table S1 for details). This sequence has one frameshift and covers 217 aa of the 284 aa used for divergence vs. age analysis in Figure 3. The conceptual translation of

this sequence has the same divergence compared to *AaegR4_1.1* as does the *AgamR4_1.1* copy, 20%. The high degree of identity extends throughout the ORFs. At the nt level, the mosquito *RTE* clade sequences have approximately 67% identity over a continuous 2.3 kb region of the ORF that could be aligned with confidence, compared to the 78% nt identity for the whole 3.6 kb ORFs of *AaegR4_1.1* and *AgamR4_1.1*. To check for additional highly conserved non-LTR retrotransposons, RT aa sequences from all identified *An. gambiae* non-LTR retrotransposon elements (Biedler and Tu, 2003) were tested as queries in TBLASTn vs. the *Ae. aegypti* genome sequence. No other non-LTR retrotransposon pairs between *Ae. aegypti* and *An. gambiae* showed such high identity as mosquito *R4* and *RTE* sequences.

Horizontal transfer can be difficult to prove when the TEs in question do not have extraordinarily high identities. There are a few alternative explanations for the observed relatively low divergence between the two mosquito *R4* clade sequences. Selection pressure could explain the high conservation between *Ae. aegypti* and *An. gambiae* elements. Purifying selection could result from selection on active elements or on co-opted elements serving a host function. High aa conservation is found throughout the length of *AaegR4_1.1* and *AgamR4_1.1* ORFs. At the nt level, there is also a high conservation throughout the full-length sequence, as well as in their 5' and 3' UTRs. The high nt conservation in the coding sequence does not appear to be influenced by codon bias. The mean effective number of codons (Wright, 1990) for *AaegR4_1.1* and *AgamR4_1.1* is 58, determined by PDA (<http://dpdb.uab.es/pda/pda.asp>) (Casillas and Barbadilla, 2004). Possible values range from 20 (high bias) to 61 (no bias). The idea that mosquito *R4* clade elements serve an important function in mosquitoes seems unlikely because copies are absent in *C. quinquefasciatus* and all Anopheline species outside of the *An. gambiae* species complex, and no intact copies can be detected in *An. gambiae*. It could be that *R4* clade elements are more highly conserved than other non-LTRs. However, the comparisons of non-mosquito *R4* clade non-LTR retrotransposons to mosquito *R4* clade representatives fit well with the regression line (points a–d in Figure 3). Therefore, the lower divergence between mosquito *R4* clade sequences does not appear to be due to properties intrinsic to the *R4* clade.

To search for other potential horizontal transfer events involving *R4* clade-like elements, we performed searches of numerous genome sequence databases. We used the *AaegR4_1* nt sequence for BLAST vs. the nucleotide collection database (NCBI) and did not detect any non-mosquito sequences. The only hit was for *DongAg*, the *R4*-like element in *An. gambiae*. We also searched 68 whole genome sequences by BLAT (nucleotide search, with e-value cutoff 1e-10) via Ensembl (ensembl.org) and found one hit in the *Anolis carolinensis* (lizard) genome. The alignment was 230 nt long, having 57% ID and an e-value of 2e-15. BLASTx of this sequence vs. the non-redundant protein database (NCBI) resulted in top hits to reverse transcriptase in *Pediculus humanis* (human louse) with the next most significant hit being *Dong*, the *R4*-like element in *Bombyx mori*. The amino acid identities for these hits are about 50%. While interesting, we do not explore the significance of these findings.

AaegR4_1 and *AgamR4_1* are the sole *R4* clade families in *Ae. aegypti* and *An. gambiae*. It appears that only *Ae. aegypti* *R4* copies have the potential to be active. Divergence vs. age analysis of the RT region suggests that the mosquito *R4* clade family was involved in a horizontal transfer event (Figure 3). We have applied the conservative estimate of 145 MYA

(145–200 MYA) for the divergence of *Ae. aegypti* and *An. gambiae* (Krzywinski et al., 2006). Use of higher divergence values will only increase the significance of our findings. Thus the most parsimonious explanation is that a mosquito *R4* clade element was horizontally transferred between aedine and anopheline lineages. A reasonable hypothesis is that transfer occurred from the aedine to the anopheline lineage and failed to thrive in *Anopheles* species. It is interesting that these species have long co-existed in parts of Africa, providing a possible ecological connection for horizontal transfer (Sinka et al., 2010, Rogers et al., 2006).

Experimental Procedures

White gene phylogeny

Amino acid sequences (S2) were obtained for the *white* gene from [Vectorbase.org](http://www.Vectorbase.org) by BLAST using the *An. gambiae white* gene as the query (Besansky et al., 1995). The *white* gene phylogeny (Figure 2) was generated using Phylogeny.fr (<http://www.phylogeny.fr>) in “one-click” mode (Dereeper et al., 2010, Dereeper et al., 2008, Edgar, 2004, Castresana, 2000, Guindon and Gascuel, 2003, Anisimova and Gascuel, 2006, Chevenet et al., 2006).

Divergence vs. age analysis

Non-LTR retrotransposon RT amino acid sequences (S3) were aligned (Thompson et al., 1997) and distance values were determined by PAUP (Swofford, 2002). Percent RT amino acid divergence was plotted vs. host divergence (MY) using Microsoft Excel and a logarithmic trend line was fitted. RStudent values were calculated using MINITAB Release 14.

PCR

PCR was performed to amplify full-length copies of *AgamR4* by PCR using genomic DNA from *An. gambiae* (PEST). Forward (5'-GAAGGCTAACCACAATAATG-3') and reverse (5'-GAATTTTGTCACTTCTCAGC-3') primers were designed using the *AgDong* sequence. PCR was performed using Takara Taq DNA Polymerase (Shimadzu Corporation) using the following thermocycler parameters: 95°C 3 min, and 35 cycles of 95°C 30 s, 53°C 30 s, 72°C 3 min 30 s.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Wang X. and Zheng H. from the Statistics Department of Virginia Tech for advice on regression analysis. We also thank Kojima K. K. and Fujiwara H. for providing details of the *DongAg* (*AgamR4_1.1*) reconstruction method. This work was supported by a NIH grant AI42121 to Z. Tu.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403–410. [PubMed: 2231712]

- Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*. 2006; 55:539–552. [PubMed: 16785212]
- Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F, Campbell CL, Campbell KS, Casola C, Castro MT, Chandramouliswaran I, Chapman SB, Christley S, Costas J, Eisenstadt E, Feschotte C, Fraser-Liggett C, Guigo R, Haas B, Hammond M, Hansson BS, Hemingway J, Hill SR, Howarth C, Ignell R, Kennedy RC, Kodira CD, Lobo NF, Mao C, Mayhew G, Michel K, Mori A, Liu N, Naveira H, Nene V, Nguyen N, Pearson MD, Pritham EJ, Puiu D, Qi Y, Ranson H, Ribeiro JM, Roberston HM, Severson DW, Shumway M, Stanke M, Strausberg RL, Sun C, Sutton G, Tu ZJ, Tubio JM, Unger MF, Vanlandingham DL, Vilella AJ, White O, White JR, Wondji CS, Wortman J, Zdobnov EM, Birren B, Christensen BM, Collins FH, Cornel A, Dimopoulos G, Hannick LI, Higgs S, Lanzaro GC, Lawson D, Lee NH, Muskavitch MA, Raikhel AS, Atkinson PW. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science*. 2010; 330:86–88. [PubMed: 20929810]
- Besansky NJ, Bedell JA, Benedict MQ, Mukabayire O, Hilfiker D, Collins FH. Cloning and characterization of the white gene from *Anopheles gambiae*. *Insect Mol Biol*. 1995; 4:217–231. [PubMed: 8825759]
- Biedler J, Tu Z. Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. *Mol Biol Evol*. 2003; 20:1811–1825. [PubMed: 12832632]
- burke WD, Muller F, Eickbush TH. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res*. 1995; 23:4628–4634. [PubMed: 8524653]
- Casillas S, Barbadilla A. PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res*. 2004; 32:W166–W169. [PubMed: 15215372]
- Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000; 17:540–552. [PubMed: 10742046]
- Chevenet F, Brun C, Banuls AL, Jacq B, Christen R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*. 2006; 7:439. [PubMed: 17032440]
- Coetzee M, Hunt RH, Wilkerson R, Della Torre A, Coulibaly MB, Besansky NJ. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa*. 2013; 3619:246–274.
- Dereeper A, Audic S, Claverie JM, Blanc G. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol Biol*. 2010; 10:8. [PubMed: 20067610]
- Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 2008; 36:W465–W469. [PubMed: 18424797]
- Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature*. 1980; 284:601–603. [PubMed: 6245369]
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; 32:1792–1797. [PubMed: 15034147]
- Eickbush DG, Eickbush TH. Vertical transmission of the retrotransposable elements R1 and R2 during the evolution of the *Drosophila melanogaster* species subgroup. *Genetics*. 1995; 139:671–684. [PubMed: 7713424]
- Eickbush, T.; Malik, H. Origins and evolution of retrotransposons. In: Craig, NL.; Gellert, RCM.; Lambowitz, AM., editors. *Mobile DNA II*. Washington, DC: American Society for Microbiology Press; 2002.
- Finnegan DJ. Transposable elements. *Curr Opin Genet Dev*. 1992; 2:861–867. [PubMed: 1335807]
- Gaunt MW, Miles MA. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol*. 2002; 19:748–761. [PubMed: 11961108]
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003; 52:696–704. [PubMed: 14530136]
- Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, DE Berardinis V, Baldwin D,

Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanagan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*. 2002; 298:129–149. [PubMed: 12364791]

Kojima KK, Fujiwara H. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Mol Biol Evol*. 2004; 21:207–217. [PubMed: 12949131]

Kojima KK, Fujiwara H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol*. 2005; 22:2157–2165. [PubMed: 16014872]

Kordis D, Gubensek F. Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. *Proc Natl Acad Sci U S A*. 1998; 95:10704–10709. [PubMed: 9724768]

Kordis D, Gubensek F. Horizontal transfer of non-LTR retrotransposons in vertebrates. *Genetica*. 1999; 107:121–128. [PubMed: 10952205]

Krzywinski J, Grushko OG, Besansky NJ. Analysis of the complete mitochondrial DNA from *Anopheles funestus*: An improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol Phylogenet Evol*. 2006; 39:417–423. [PubMed: 16473530]

Lathe WC 3rd, Burke WD, Eickbush DG, Eickbush TH. Evolutionary stability of the R1 retrotransposable element in the genus *Drosophila*. *Mol Biol Evol*. 1995; 12:1094–1105. [PubMed: 8524042]

Lathe WC 3rd, Eickbush TH. A single lineage of r2 retrotransposable elements is an active, evolutionarily stable component of the *Drosophila* rDNA locus. *Mol Biol Evol*. 1997; 14:1232–1241. [PubMed: 9402733]

Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*. 1999; 16:793–805. [PubMed: 10368957]

Marinotti O, Cerqueira GC, De Almeida LG, Ferro MI, Loreto EL, Zaha A, Teixeira SM, Wespiser AR, Almeida ESA, Schlindwein AD, Pacheco AC, Silva AL, Graveley BR, Walenz BP, Lima Bde A, Ribeiro CA, Nunes-silva CG, De Carvalho CR, Soares CM, De Menezes CB, Matioli C, Caffrey D, Araujo DA, De Oliveira DM, Golenbock D, Grisard EC, Fantinatti-Garboggini F, De Carvalho FM, Barcellos FG, Prosdocimi F, May G, Azevedo Junior GM, Guimaraes GM, Goldman GH, Padilha IQ, Batista Jda S, Ferro JA, Ribeiro JM, Fietto JL, Dabbas KM, Cerdeira L, Agnez-Lima LF, Brocchi M, De Carvalho MO, Teixeira Mde M, Diniz Maia Mde M, Goldman MH, Cruz Schneider MP, Felipe MS, Hungria M, Nicolas MF, Pereira M, Montes MA, Cantao ME, Vincentz M, Rafael MS, Silverman N, Stoco PH, Souza RC, Vicentini R, Gazzinelli RT, Neves Rde O, Silva R, Astolfi-Filho S, Maciel TE, Urmenyi TP, Tadei WP, Camargo EP, De Vasconcelos AT. The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Res*. 2013; 41:7387–7400. [PubMed: 23761445]

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyne B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorri H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanze J, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW. Genome

- sequence of *Aedes aegypti*, a major arbovirus vector. *Science*. 2007; 316:1718–1723. [PubMed: 17510324]
- Novikova O, Fet V, Blinov A. Non-LTR retrotransposons in fungi. *Funct Integr Genomics*. 2009; 9:27–42. [PubMed: 18677522]
- Novikova O, Sliwinska E, Fet V, Settele J, Blinov A, Woyciechowski M. CR1 clade of non-LTR retrotransposons from *Maculinea* butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission. *BMC Evol Biol*. 2007; 7:93. [PubMed: 17588269]
- Rogers DJ, Wilson AJ, Hay SI, Graham AJ. The global distribution of yellow fever and dengue. *Adv Parasitol*. 2006; 62:181–220. [PubMed: 16647971]
- Sanchez-Gracia A, Maside X, Charlesworth B. High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet*. 2005; 21:200–203. [PubMed: 15797612]
- Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol*. 2010; 25:537–546. [PubMed: 20591532]
- Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, Patil AP, Temperley WH, Gething PW, Kabaria CW, Okara RM, Van Boeckel T, Godfray HC, Harbach RE, Hay SI. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasit Vectors*. 2010; 3:117. [PubMed: 21129198]
- Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O. Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in *Lepidoptera* species. *Mol Biol Evol*. 2012; 29:3685–3702. [PubMed: 22826456]
- Swofford, DL. Version 4.0b10. Sunderland, Massachusetts: Sinauer Associates; 2002. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997; 25:4876–4882. [PubMed: 9396791]
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994; 22:4673–4680. [PubMed: 7984417]
- Volff JN, Korting C, Froschauer A, Sweeney K, Scharf M. Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. *J Mol Evol*. 2001; 52:351–360. [PubMed: 11343131]
- Volff JN, Korting C, Scharf M. Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. *Mol Biol Evol*. 2000; 17:1673–1684. [PubMed: 11070055]
- Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990; 87:23–29. [PubMed: 2110097]
- Xiong Y, Eickbush TH. Dong, a non-long terminal repeat (non-LTR) retrotransposable element from *Bombyx mori*. *Nucleic Acids Res*. 1993; 21:1318. [PubMed: 8385316]
- Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, Xu Y, Zhou H, Xiong C, Li S, Yu J, Hong S, Yu X, Zou P, Chen C, Chang X, Wang W, Lv Y, Sun Y, Ma L, Shen B, Zhu C. Genome sequence of *Anopheles sinensis* provides insight into genetics basis of mosquito competence for malaria parasites. *BMC Genomics*. 2014; 15:42. [PubMed: 24438588]
- Zupunski V, Gubensek F, Kordis D. Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. *Mol Biol Evol*. 2001; 18:1849–1863. [PubMed: 11557792]

Figure 1. Schematic of *AaegR4_I* structural organization
AaegR4_I.1 is approximately 4.0 kb and contains a single 3.6 kb ORF encoding reverse transcriptase (RT) and a restriction enzyme-like endonuclease (REL). Full-length copies are flanked by TAA tandem repeats. The location of a CCHC nucleic acid-binding motif is indicated.

Figure 2. Host species phylogeny and mosquito *R4* clade sequence distribution

Mosquito phylogeny including 7 species of 3 genera based on the *white* gene. Shaded boxes show species containing *R4* clade sequences. *Ae. aeg.* (*Ae. aegypti*); *Ae. alb.* (*Ae. albopictus*); *C. qui.* (*C. quinquefasciatus*); *A. dar.* (*An. darlingi*); *An. sin.* (*A. sinensis*); *A. ste.* (*An. stephensi*); *A. gam.* (*An. gambiae*). Divergence time between aedine and anopheline lineages of 145 to 200 Million Years Ago (MYA) is from Krzywinski et al., 2006. Maximum likelihood phylogeny was constructed using Phylogeny.fr online at http://www.phylogeny.fr/version2_cgi/simple_phylogeny.cgi. Scale bar shows substitutions per site. Branch support values are indicated above branches as percentages. Sequences used for phylogeny can be found in S2.

Figure 3. RT divergence vs. age of host divergence

Non-LTR retrotransposon RT sequences were aligned (Thompson et al., 1997) and distance values were determined by PAUP (Swofford, 2002). Points 1–10 represent arthropod comparisons taken from Malik, Burke, and Eickbush, 1999. Points a and b represent *B. mori* Dong compared to *AgamR4_1.1* (a) and *AaegR4_1.1* (b). Points c and d represent *A. lumbricoides* R4 compared to *AgamR4_1.1* (c) and *AaegR4_1.1* (d). Dipteran (mosquito)/Lepidopteran divergence is estimated at 340 MY (Gaunt and Miles, 2002), and 600 MY is used for *A. lumbricoides*/mosquito divergence. Divergence times are taken from Malik et al., 1999. Thirteen mosquito non-LTR retrotransposon comparisons representing all 10 non-LTR retrotransposon clades in *An. gambiae* (e–q) are shown in the rectangle where 145 MYA was used as a conservative estimate for *Ae. aegypti* and *An. gambiae* divergence. A logarithmic trendline is shown and statistics were calculated using log host age vs. amino acid divergence. Pearson's correlation coefficient (r) is given with p -value. RStudent, Studentized deleted residual. An asterisk indicates a value based on the inclusion of only one of the 13 *Ae. aegypti*/*An. gambiae* non-LTR retrotransposon comparisons, the mosquito R4 clade elements (RStudent = 5.7). This removes the influence of other mosquito comparisons that may include paralogous sequences or sequences that have unusually high conservation. All sequences and distance values are available in S3.