Transposable element dynamics of the hAT element *Herves* in the human malaria vector, *Anopheles gambiae s.s.*

Ramanand A. Subramanian,* Peter Arensburger,† Peter W. Atkinson,† David A. O’Brochta*¹

*Center for Biosystems Research; University of Maryland Biotechnology Institute; 9600 Gudelsky Drive, Rockville, Maryland, 20850, †Department of Entomology, University of California, Riverside, California, 92521-0314

*Herves*, transposable elements, *hAT* elements, *Anopheles gambiae*, malaria

¹Corresponding author: Center for Biosystems Research, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850.

E-mail: obrochta@umbi.umd.edu

Phone: 240-314-6343

Fax: 240-314-6255
ABSTRACT
Transposable elements are being considered as genetic drive agents for introducing phenotype-altering genes into populations of vectors of human disease. Transposable element dynamics of endogenous elements will assist in predicting the behavior of introduced elements. Transposable element display was used to estimate the site occupancy frequency distribution of *Herves* in six populations of *Anopheles gambiae s.s.* The site occupancy distribution data suggest that the element has been recently active within the sampled populations. All 218 individuals sampled contained at least one copy of *Herves* with a mean of 3.6 elements per diploid genome. No significant differences in copy number were observed among populations. Nucleotide polymorphism within the element was high (\( \pi = 0.0079 \) in non-coding sequences and 0.0046 in coding sequences) relative to that observed in some of the more well-studied elements in *D. melanogaster*. In total, 33 distinct forms of *Herves* were found based on the sequence of the first 528 bp of the transposase open reading frame. Only 2 forms were found in all six study-populations. Although *Herves* elements in *An. gambiae* are quite diverse, 85% of the individuals examined had evidence of complete forms of the element. Evidence was found for the lateral transfer of *Herves* from an unknown source into the *An. gambiae* lineage prior to the diversification of the *An. gambiae* species complex. The characteristics of *Herves* in *An. gambiae* are somewhat unlike those of *P* elements in *D. melanogaster.*
INTRODUCTION

*hAT* elements comprise a large and prevalent group of Class II transposable elements found in a wide range of plants, animals (KEMPKEN and WINDHOFER 2001; KUNZE and WEIL 2002; RAY et al. 2007). *hAT* elements are not only of interest for their role in genome evolution but also as tools for genetically modifying organisms, with the elements *Hermes* and *hobo* being two examples of *hAT* element-derived insect gene vectors (BLACKMAN et al. 1989; O'BROCHTA et al. 1996). Transposable elements from other families such as *piggyBac, Mos I* and *Minos* have also been developed into effective insect gene vectors that are now employed in a variety of applications (ATKINSON et al. 2001). Using these relatively new gene-integration tools, a novel form of biological control is being considered to stem the transmission of certain arboviruses (e.g. Dengue) and parasites (e.g. *Plasmodium*) by mosquitoes and other arthropod vectors (ADELMAN et al. 2002; ALPHEY et al. 2002; BEARD et al. 2002). This strategy involves the introduction of transgenic insects into natural populations of a target species with the intent of replacing the native population with genetically modified con-specifics (ANONYMOUS 1991; CRAIG 1963; JAMES 1992; MILLER 1992). Introduced transgenic mosquitoes will contain transgenes conferring incompatibility (refractoriness) or resistance to the target pathogen or parasite. An increase in the frequency of the transgene within natural populations of the vector will, under certain conditions, lead to a reduction or elimination of vector-borne disease transmission (BOETE and KOELLA 2002).

Designing gene vectors and effector transgenes for refractoriness such that they will increase in natural populations and eventually reach fixation is a considerable challenge and transposable elements may provide a means by which this can be accomplished (BRAIG and YAN 2001). The replicative nature of transposable element movement (even by elements that move
by a cut-and-past fashion ie. Class II elements) results in elements acquiring a transmission advantage, resulting in their gradual increase in frequency in populations (KISZEWSKI and SPIELMAN 1998; RIBEIRO and KIDWELL 1994). The magnitude of that transmission advantage is determined by the rate of transposition, the degree to which transposition is conservative or replicative, the spatial patterns of element transposition within a genome, the biology of the transposable element and its interactions with the host insect, and the size, structure and characteristics of the target population (RASGON and GOULD 2005).

While intra-species spreading of transposable elements through transposition has been observed in nature following recent horizontal transfer events involving transposable elements (e.g. \textit{P} and \textit{hobo} elements), population modification has never been attempted by the deliberate and intentional release of an active autonomous transposable element into natural populations of insects (ROBERTSON 2002). Predicting the outcome of such an intentional release of transgenic mosquitoes containing active autonomous transposable element gene vectors is an enormous challenge but one that must be successfully met if population replacement biological control using transposable elements is to be successful (ALPHEY \textit{et al.} 2002). Data that might inform those predictions include an understanding of the dynamics of endogenous Class II transposable elements within the host insect. Endogenous elements are likely to reveal temporal and spatial patterns of spread as well as how population structure has influenced those patterns. Currently our understanding of the population dynamics of Class II transposable elements in insects is based almost entirely on studies of \textit{P} and \textit{hobo} elements in \textit{D. melanogaster} and closely related species (ANXOLABEHERE \textit{et al.} 1990; ANXOLABEHERE \textit{et al.} 1988; BUCHETON \textit{et al.} 1992; SILVA and KIDWELL 2004; SIMMONS 1992). These studies have documented the ability of these elements to spread rapidly through populations and for the elements to become structurally
modified over time, most often by internal deletion. The propensity of these elements to accumulate internal deletions rapidly has raised a serious concern about using transposable elements as transgene spreading agents, namely, the frequent loss of transgenes. Maintaining tight linkage between the anti-parasite effector gene and the associated gene drive system has been repeatedly stated as an essential characteristic of this biological control strategy (CURTIS 2003; JAMES 2005). To what extent these characteristics of \( P \), \( hobo \) and \( mariner \) elements are general characteristics of Class II elements remains to be fully explored. Because a proposed target species for this novel population replacement-based biological control strategy is the human malaria vector, \textit{Anopheles gambiae}, the study of Class II transposable element dynamics in this species is particularly relevant.

Recently, a functional \textit{hAT} element, \textit{Herves}, was discovered in \textit{An. gambiae}, providing an opportunity to examine the dynamics of an active Class II transposable element in this insect (ARENSBURGER \textit{et al.} 2005). \textit{Herves} is notably different at the sequence level from the well-studied \textit{hobo} element from \textit{D. melanogaster} and \textit{Hermes} from \textit{Musca domestica}, sharing only about 20% amino acid identity with these elements (ARENSBURGER \textit{et al.} 2005). A \textit{Herves} element isolated from the RSP strain of \textit{An. gambiae} that was established as a laboratory colony in the early 1990s (VULULE \textit{et al.} 1994) was shown to be transpositionally active in laboratory-based mobility assays in \textit{D. melanogaster} (ARENSBURGER \textit{et al.} 2005) and \textit{Aedes aegypti} (P. Arensburger and P. Atkinson, unpublished). A recent study of the element’s abundance and site-occupancy frequency in natural populations of \textit{An. gambiae} s.s., \textit{An. merus}, and \textit{An. arabiensis} in Mozambique revealed that it was present in all three species at approximately 5 copies per diploid genome and site-occupancy frequency distributions suggested that \textit{Herves} had been recently active in the three species examined (O’BROCHTA \textit{et al.} 2006). In the population of \textit{An.}
An. gambiae examined in Mozambique, 95% of the individuals tested contained intact (non-deleted) forms of the element which is quite unlike $P$ elements in *D. melanogaster* in which most elements are internally deleted derivatives of the canonical element (O'HARE et al. 1992). Here *Herves* has been investigated in six populations of An. gambiae using a variety of methods to see if the characteristics of the element observed in Mozambique were general features of the element and how it compares to other well-studied Class II elements.
MATERIAL AND METHODS.

Collection Site: *Anopheles gambiae* s.s. from six populations were used in this study with sample sizes ranging from 15-94 individuals (Figure 1). Samples from Asembo Bay (hereafter referred to as Asembo), Kisian and Malindi have been described (LEHMANN et al. 2003). Asembo and Kisian are located in western Kenya and were sampled in 1994 and 1996 respectively (LEHMANN et al. 2003). Malindi, located in eastern Kenya, was sampled in 1996 (LEHMANN et al. 2003). The northeastern region of Tanzania was sampled in 2004 in the region in and around the village of Zenet (MEERAUS et al. 2005). Samples from southern Mozambique (Furvela) were collected in 2003 as described (O'BROCHTA et al. 2006). Samples from north-central Nigeria (Bakin Kogi) were collected in 1999 (LEHMANN et al. 2003).

DNA Isolation: Genomic DNA was isolated from individual mosquitoes as described (O'BROCHTA et al. 2006) and resuspended in 100 µl of distilled water and stored at -80°C.

Species Identification: Species identification was performed using the method of Scott et al. (1993) as described (O'BROCHTA et al. 2006) using 1/100th of the total genomic DNA from a single mosquito in a volume of 1µl (SCOTT et al. 1993). This method permits the identification of species-specific polymorphisms in the intergenic spacer region of ribosomal RNA genes using PCR. Only *An. gambiae* s.s. samples yielding unambiguous species identification results were used in subsequent analyses.

Transposable element display: Transposable element display is a PCR-based DNA fingerprinting method derived from the Amplified Fragment Length Polymorphism (AFLP) method (VOS et al. 1995). It was performed here as described previously with only minor modifications (O'BROCHTA et al. 2006). Transposable element display was performed in triplicate using 2-5 µl (approximately 200ng) of genomic DNA for each replicate. Genomic
DNA was digested for 4 hours in a volume of 40 µl at 37°C with 4 units of the restriction endonuclease *MseI* using conditions recommended by the manufacturer (New England Biolabs). Sixty picomoles of adapters were ligated to the *MseI* digestion products by adding 10 µl of 1X restriction enzyme buffer containing 5 mM ATP, 50 mM DTT (dithiothreitol), 10 µg BSA (bovine serum albumin), 4 units of *MseI*, 1 Weiss unit of T4 DNA ligase and incubated at 37°C overnight. The adapters were prepared by mixing equimolar amounts of oligonucleotides HhaIa (5' GAT GAG TCC TGA GTA CG 3') and MseIb2 (5' TAC GTA CTC AGG ACT CAT CAA G 3'), heating them to 100°C for 10 minutes and then allowing the mixture to very slowly cool to room temperature. The design of the adapters and the digestion/ligation reaction conditions result in the efficient creation of only monomeric *MseI*-cut genomic DNA fragments with terminal adapters.

Five microliters of the restriction/ligation reaction were used as the template in a polymerase chain reaction (“preselective PCR”) performed in a 50 µl reaction volume containing 1X PCR Buffer II (Applied Biosystems), 0.2 mM dNTPs (an equimolar mixture of dATP, dTTP, dCTP, dGTP), 2.5 mM MgCl₂, 1 unit AmpliTaq® DNA polymerase (Applied Biosystems), and 24 pmoles of primer HhaIa and primer HervTEDAL1a (5' ATT TCG ACG GGT TCC TAC C 3'). HervTEDAL1a is a *Herves*-specific primer that anneals to sequences approximately 150 bp from the 5’ end of the element. The DNA polymerase was added as a complex with TaqStart™ Antibody (Clontech) as described by the manufacturer for the purposes of “hot-starting” the reaction. The reaction conditions were 95°C/3 mins followed by 25 cycles of 95°C/15 sec, 63°C/30 sec, 72°C/1.0 min and a final cycle of 72°C/5 min. A second PCR was performed (“selective PCR”) using 5 µl of the preselective PCR products as template in a 20 µl reaction containing 1X PCR Buffer II, 0.2 mM dNTPs, 2.5 mM MgCl₂, 1 unit AmpliTaq® DNA
Polymerase (bound to TaqStart™ Antibody as above), 9 pmoles each of primers HhaIa and Cy5™-labeled HervTEDAL2 (5’ GTT GAT TAG ATG AAC GTA GG 3’). The Cy5™-labeled primers were purified by HPLC prior to their use. HervTEDAL2 anneals to sequences approximately 80 bp from the 5’ end of the element. Following a denaturation step at 95°C for 3 minutes “touchdown” PCR conditions were created in which during the first 5 cycles the annealing temperature was decreased 1°C after each cycle with the first of these cycles being 95°C/15 sec, 64°C/30 sec, 72°C/1.0 min. Following these 5 cycles an additional 25 cycles were performed at 95°C/15 sec, 60°C/30 sec, 72°C/1.0 min with a final cycle of 72°C/5 min.

To visualize products of transposable element display 5 µl of selective PCR products were mixed with 5µl of loading buffer (95% deionized formamide, 10mM EDTA) heated to 95°C for 3 minutes, cooled quickly on ice and 6 µl were loaded on a 6% polyacrylamide gel (19:1 acrylamide : bisacrylamide) containing 6.7 M urea in 1X TBE buffer (90 mM Tris-borate, 2 mM EDTA). ALFExpress™Sizer™50-500 (Amersham/Pharmacia) was used as a size standard. Electrophoresis was performed at 70 watts (constant) for 2.5 hours at which time the gel was transferred to 3MM filter paper and dried. The dried gel was scanned on a STORM 860 phosphoimager (Molecular Dynamics). The products obtained from the three independent replicate reactions of the same sample were run on the same gel to assist with determining the presence of bands. Based on the combined results of three transposable element display experiments a band was called as present or absent if it was unambiguously present in at least 2 of the 3 replicates. Determining the presence of bands in this way resulted in a single scoring matrix that was then used in subsequent analyses.

Site-occupancy frequency distributions were estimated using transposable element display data. Using the frequency distributions and assuming the model of Charlesworth and
Charlesworth (1983) the model parameter $\beta$, that measured, in part, the forces removing insertions from natural populations, was estimated. The model parameter $\beta$ is equal to the product of four times the effective population size and the rate of element loss. Estimation of $\beta$ and the copy number of *Herves* per diploid genome were performed as described by Wright *et al.* (2001) who considered the dominant nature of transposable element display signals and the application of the parameter estimation methods of Charlesworth and Charlesworth (1983) to diploid organisms. Note that although each sample was analyzed three times for transposable element display these replicates were used to produce a single scoring matrix. The advantage of this procedure is that it increased the accuracy of determining the presence of bands and minimized errors that tend to result in overestimations of $\beta$.

**Transposase Open Reading Frame Detection:** To assess *Herves* open reading frames for the presence of deletions and insertions, PCR primers were designed that were complementary to sequences flanking the transposase ORF: 1372f (5′ CCA CAA ATT GAT CTA CGC TCC 3′) and 3469r (5′ GAT GCA TCT ATT ATG ATT AAG GC 3′). One fiftieth of the genomic DNA from one mosquito (2 µl) was used as template in a 50µl reaction containing 1X ThermalAce™ (Invitrogen), 0.2 mM dNTPs (an equimolar mixture of dATP, dTTP, dCTP, dGTP), 2.5 mM MgCl$_2$, 2 units ThermalAce™ DNA polymerase (Invitrogen), and 24 pmoles of primer1372f and 3469r. Amplification reactions were performed under the following conditions: 95°C/3 min followed by 30 cycles of 95°C/30 sec, 48°C/30 sec, 72°C/3.0 min and a final cycle of 72°C/10 min. Reaction products were fractionated on a 1% agarose gel. PCR products of the samples that failed to produce a detectable product following one round of PCR were used as templates (5µl) in a second PCR under the same conditions described above but with primers 1407f (5′ GAT CAA AGG TAA CAT TAG TCT TG 3′) and 3294r (5′ CCA
TGT TAC AAA TTT TGC AAC G 3’) and rechecked on a 1% agarose gel. Open reading frames free of deletions and insertions yielded PCR products 2100 bp after the first PCR and 1900 bp after the second PCR. We estimate that elements with deletions as small as 100 bp would be detectable using this strategy.

Sampling and PCR for population analysis: Transposable element display permitted occupied sites to be identified and these data were used in determining the composition of the subset of individual mosquito genomic DNAs that would be used in the analysis of sequence diversity of 1474 bp of the non-coding region and the first 528 bp of the transposase open reading frame. This selected subset of individual mosquito genomic DNAs was such that Herves elements at most occupied sites, as determined by transposable element display, were included in the PCR template pool. So, a total of 49 individuals containing elements at the 130 different sites identified by transposable element display were included in the PCR template pool to give us an opportunity to amplify Herves elements inserted at different genomic sites within the populations. Using this subset of genomic DNAs a portion of the left end of the element was amplified using a nested PCR strategy. Five microliters of genomic DNA from each of the 49 individuals were used as template in separate 20µl reactions containing 5X Phusion HF Buffer (NEB), 0.2µM dNTPs (an equimolar mixture of dATP, dTTP, dCTP, dGTP), 0.4 units Finnzymes Phusion™ DNA polymerase (New England Biolabs; error rate = 4.4 x 10⁻⁷), and 24 pmoles of primer 24F (5’ TAG AGT TGT GCC TCA AGA ACC AGA 3’) and primer 2035R (5’ TGG TTC AGG TTT GTC CAT CC 3’). Amplification reactions were preformed under the following conditions: 98°C/1 min followed by 25 cycles of 98°C/10 sec, 65°C/30 sec, 72°C/1 min 30 sec and a final cycle of 72°C/10 min. Reaction products were fractionated in a 1% agarose gel. PCR products from samples that failed to produce detectable products on an agarose gel
following one round of PCR were used as templates (5µl) in a second PCR under the same conditions described above using primers 24F (5’ TAG AGT TGT GCC TCA AGA ACC AGA 3’) and 2002r (5’GCT ATA GCT TTG GCG GTC G 3’) and rechecked on a 1% agarose gel. The 2kb amplification product was eluted from the gel, precipitated, resuspended in 20 µl dH2O and cloned into the pCR®-Blunt II TOPO vector (Invitrogen). Up to five clones per individual were sequenced and these sequences were used in subsequent analyses. From samples “Zenet”, “Asembo”, “Bakin-Kogi”, “Kisian”, “Furvela” and “Malindi” a total of 57 (GenBank accessions EF588609-EF588665), 51 (EF588428-EF588478), 40 (EF588479-EF588518), 29 (EF588552-EF588580), 33 (EF588519-EF588551) and 28 (EF588581-EF588608) sequences, respectively, were obtained. Note, the methods used to obtain the sequences for this analysis did not permit these elements to be assigned to specific sites identified in the site-occupancy (transposable element display) analysis.

**Sequence Analysis:** Sequences were aligned using AlignX, a ClustalW-base alignment program in VectorNTI Advance 10.0.1 (Invitrogen). Nucleotide diversity was estimated from average pair-wise number of differences between elements, $\pi$ (Nei and Li 1979) and from the number of polymorphic sites, $\theta$ (Waterson 1975). $\pi$ and $\theta$ were estimated using DnaSP 3 (Rozas and Rozas 1995; Rozas et al. 2003). Estimates of the observed silent site diversity in the first 528 bp of the 5’ end of the transposase coding region was computed using the Kumar method (Nei and Kumar 2000) as implemented in MEGA 3.1 (Kumar et al. 2004). Expected values of silent site diversity were calculated following Sanchez-Garcia et al. (2005) and was the product of the haploid copy number and the average synonymous diversity (0.0209) from a sample of 35 nuclear genes (Morlaís et al. 2004). Tajima’s D was calculated using DnaSP 3. Further analysis was performed on the first 528 bp of the 5’ end of the transposase open reading
frame. Unique variants of elements were identified (referred to as forms), their frequencies determined and the relationship of the forms determined using TCS1.21 (CLEMENT et al. 2000). Alignment gaps were treated as missing data in this analysis. Estimates of the number of synonymous substitutions per synonymous site (dS) and of non-synonymous substitutions per non-synonymous site (dN) and their ratio, $\omega = dN/dS$, were obtained using maximum likelihood (ML) methods employed by CODEML in PAML 3.13 (YANG 1997) using the alignment of the 33 different forms for the analysis (Supplemental Figure 1). PAML permits an assessment of the observed substitution data after assuming different codon substitution models that differ in the way selection pressure is distributed within the gene. Here we have examined our data in light of three simple models: a single ratio model (M0) that assumes one $\omega$ for all sites, a neutral model (M1) that assumes that there are two classes of sites within the gene; those that are conserved ($p_0$) with $\omega_0=0$ and those that are neutral ($p_1=1-p_0$) with $\omega_1=1$, and finally, a discrete model (M3) that assumes three classes of sites each having a unique value of $\omega$ that is estimated from the data. A likelihood ratio test (LRT), which is twice the log-likelihood difference between two models being compared, was used to determine which model best reflected the observed data. The LRT statistic has a $X^2$ distribution with degrees of freedom equal to the difference in the number of parameters between the two models (YANG et al. 2000).
RESULTS

**Site Occupancy:** Transposable element display has been a useful tool for assessing the number and position of transposable elements within the genome of individual organisms (Biedler et al. 2003; Guimond et al. 2003; Wright et al. 2001). As performed here, templates longer than 1 kb are likely to be poorly represented because the length of the extension reactions during PCR was only one minute. Because the *An. gambiae* genome is composed of 64.8% adenines and thymines and we produced PCR templates by digesting the genomic DNA with *Mse*I (TTAA) we expected only 0.004% of the resulting fragments to be 1 kb or more in length. (We estimated this by determining what percent of the fragments greater than 80 bp were over 1kb in length. Eighty base pairs is the invariable amount of *Herves* DNA contained in each PCR product. We assumed fragment sizes following *Mse*I digestion would have an exponential distribution with $\lambda = 0.324^4$. Therefore, 0.415 of all fragments were greater than 80 bp and 0.0017 of all fragments were greater than 1 kb.) Consequently, few elements will be undetected because they are on excessively long templates. Restriction site polymorphism can result in increased estimates of site occupancy diversity since an element at one site would be displayed as two bands of different lengths resulting in those bands being scored as two elements occupying two sites. While restriction site polymorphism will have this affect on the analysis, the frequency of such polymorphisms is expected to be very low based on the known level of nucleotide polymorphism in *An. gambiae s.s.* (Morlaïs et al. 2004) and our failure to detect the same element in two different positions in transposable element displays following band isolation, reamplification and DNA sequencing (Guimond et al. 2003) and (R. A. Subramanian and D. O’Brochta,
unpublished). Confounding effects of restriction site polymorphism will be small and are not a significant source of variation in transposable element display.

All individuals in this study that were analyzed by transposable element display (n = 218) contained at least one *Herves* element (Table 1). Element copy numbers within the six populations analyzed ranged from 2.9-4.4 elements per diploid genome as calculated using the method of Wright et al (2001). No individuals were found in any population that contained more than 7.0 elements. In all populations there was an abundance of occupied sites that were observed in only small numbers of individuals (Figure 2). In Zenet, Malindi and Furvela elements with high site occupancy frequencies were observed although none of these elements were shared (Figure 2).

Charlesworth and Charlesworth (1983) and Langley et al (1983) provided theoretical frameworks for understanding site occupancy frequency distributions which could also be used to estimate element mobility rates under certain conditions. Both models can be expressed using a single parameter (\( \beta \)), assume that the elements are at equilibrium and that there are an infinite number of insertion sites within the genome. According to the models (CHARLESWORTH and CHARLESWORTH 1983; LANGLEY et al. 1983) parameter values greater than one indicate the existence of forces other than drift (mobility and/or selection) that have played a major role in shaping the observed distribution. In this study estimates of \( \beta \) were, in all cases, greater than one suggesting that element mobility played a significant role in shaping the observed distribution (Table 1).

**Nucleotide Polymorphism:** Approximately 2 kb of sequence beginning at the left (5’) inverted terminal repeat and through the first 528 bp of the transposase open reading frame was amplified,
cloned and sequenced (Figure 3). A total of 238 sequences containing the first 528 bp of the transposase open reading frame were analyzed from six different locations. The average nucleotide polymorphism in the 1474 bp of non-coding sequence ($\pi = 0.0079$) was significantly different from the polymorphism observed in the coding region ($\pi = 0.0046$; P < 0.001) (Table 2). Within the non-coding region the observed polymorphisms were non-uniformly distributed in a 666 bp region beginning at nucleotide 568 having a highly reduced level of polymorphism (Figure 3). This region corresponds to a large stretch of DNA with unknown function 5’ of the transposase-coding region and just 3’ of a pair of 100 bp sub-terminal tandem repeats (ARENDBURGER et al. 2005).

Levels of silent site diversity in *Herves* elements were compared to the average silent site diversity for single-copy host genes (Table 3) as part of an effort to look for evidence of lateral introduction of *Herves* into the *An. gambiae* lineage (BROOKFIELD 1986; SANCHEZ-GRACIA et al. 2005). The observed levels of silent diversity among *Herves* elements ranged from 3 to 125-fold less than the silent site diversity seen on average in 35 nuclear genes (MORLAIS et al. 2004). In addition, Tajima’s D statistic was calculated and found to be insignificant for each location although when calculated based on the pooled data it was significant (1.91; P<0.05; Table 3) indicating an excess of low frequency variants (TAJIMA 1989).

**Structural Integrity:** Class II transposable elements can be autonomous or non-autonomous. Autonomous elements encode for functional transposase and can undergo transposition. Non-autonomous elements can not encode for functional transposase usually as a result of deletions that remove some or all of the coding region. *P* elements in *Drosophila*, for example, often exist in forms that contain large deletions of internal sequences leaving only terminal and sub-terminal sequences resulting in non-autonomous elements (ENGELS 1989). The complete *Herves* open
reading frame is approximately 1.8 kb in length and the structural integrity of *Herves* elements was assessed by amplifying this region using primers flanking this region. *Herves* elements without any deletions resulted in PCR products of 2 kb in length and elements with deletions 100 bp or more produced distinct products less than 2 kb. Of the 218 individuals tested from six locations 85% showed evidence of the presence of complete open reading frames (Table 4). Individuals with complete elements were least abundant in Nigeria (Bakin Kogi) where only 44% showed evidence of complete open reading frames (*N* = 32). In western Kenya intact forms of the element were found in 100% of the individuals from Asembo (*N* = 24) and 90% of the individuals from Kisian (*N* = 15). In eastern coastal Kenya (Malindi, *N* = 25) and northeastern coastal Tanzania (Zenet, *N* = 73) approximately 85% of the individuals tested contained intact forms of the element. In southern Mozambique (Furvela, *N* = 49) 95% of the individuals sampled contained intact elements.

**Genealogical Relationships:** A genealogical analysis of the *Herves* elements, based on the first 528 bp of coding sequence, was performed and resulted in the identification of 33 forms among the 238 sequences that were analyzed (Table 5, Figure 4). Form-diversity (the equivalent of haplotype diversity and measured using the same algorithm) varied among locations and ranged from a low of 0.565 in Bakin Kogi to a high of 0.903 in Zenet (Table 5). Of the 33 forms, only 2 (Form 1 and Form 2) were found at all six sampling locations (Figures 4 and 5) and these comprised 51% (*n* = 238) of the elements analyzed. Twenty-four forms were found at only single locations (Figure 5, Table 6). Form 2 was the most abundant form in Bakin Kogi, Asembo, Malindi and Kisian (Figure 1). In northeastern Tanzania (Zenet) where form-diversity was highest the most abundant form was Form 5, a form that is closely related to Form 2 (Figure 4). In southern Mozambique (Furvela) however, a unique form (Form 30) was most abundant.
and comprised 21% of the 57 sequences analyzed from this location. Form 30 was highly
diverged from the abundant Form 2 and consequently was one of the most unusual elements
encountered in this analysis; only Form 31 and Form 32 from Zenet were more divergent (Figure
4). Zenet was unusual among the locations analyzed because it had the greatest number of forms
(17), 10 of which were unique to this location. Not only were there a large number of element
forms at this location but also the diversity of elements was very high. On average each location
had 9.67 forms (±4.27) and shared 3.6 forms (±1.4) with other locations (Table 6).

**Natural Selection:** We tested for evidence of selective constraints within the transposase open
reading frame by estimating $\omega$ (the ratio $d_{S}/d_{S}$) using maximum likelihood. The $\omega$ ratios ranged
from 0.41-0.71 under all models (M0, M1 and M3; see Material and Methods) revealing
evidence of purifying selection. The neutral model (M1) was rejected when compared to the
discrete model (M3) that allows for 3 classes of sites with different values of $\omega$. The LRT
statistic, $2\Delta l$ ($2\Delta l = 2(-1037.77 - (-1028.00))$, for this comparison was 19.54, which was greater
than the critical value of $X^2_{[0.001,2]} = 13.816$.

**DISCUSSION**

Understanding the dynamics of active transposable elements in *An. gambiae* will inform
predictions concerning the outcomes of biological control efforts by population replacement
using transposable elements as gene drive agents. While there have been studies that have
looked at the evolutionary history of Class II transposable elements in insects, few studies
involving insects other than *Drosophila* have attempted to examine the dynamics of Class II
transposable elements at the population level, making the current study of *Herves* in *An. gambiae* somewhat unique.

Here we examined the dynamics of *Herves* by measuring the site-occupancy frequency, nucleotide-sequence diversity and by performing a genealogical analysis of the element. The rare occurrence of locally fixed, *Herves*-occupied sites and the widespread abundance of sites that are occupied in only a few individuals are consistent with there being recent activity of *Herves* within *An. gambiae*. The site-occupancy levels observed in this study ($\beta_{\text{Herves}} = 1.9-11.0$) were similar or somewhat lower than those reported for putatively active transposable elements in *D. melanogaster*: $\beta_{\text{P element}} = 16.6$ (AJIOKA and EANES 1989), $\beta_{\text{P element}} = 5.85$ (BIEMONT et al. 1994), $\beta_{\text{copia}} = 9.79$ (BIEMONT et al. 1994), $\beta_{\text{copia}} = 16.9$ (LEIGH-BROWN and MOSS 1987), $\beta_{\text{copia}} = 48.3$ (KAPLAN and BROOKFIELD 1983).

*An. gambiae* is distributed almost continuously throughout its range in Africa and demes are likely to be large and diffuse (LEHMANN et al. 1998). Little population differentiation between populations separated by up to 50 km has been reported (LEHMANN et al. 1997) and this has also been found over distances of 6000 km (LEHMANN et al. 1996). Lehmann et al. (1998) suggest that Wright’s isolation by distance model may best describe the relationships among populations (WRIGHT 1951). Population admixture might be contributing to the pattern of site-occupancy observed in this study. However, consistent with the idea that *Herves* is currently capable of transposing in natural populations of *An. gambiae* is the finding that *Herves* elements isolated from *An. gambiae* collected from the field within the last 20 years are active when introduced into other insects in the laboratory (ARENSBURGER et al. 2005).

A number of pieces of data indicate that *Herves* entered the *An. gambiae* lineage via a horizontal gene transfer. First, a comparison of the silent site diversity among *Herves* elements
and 35 nuclear genes (MORLAIS et al. 2004) revealed less diversity within *Herves* transposable elements than expected assuming similar mutation rates apply to Class II transposable elements and nuclear genes (SANCHEZ-GRACIA et al. 2005). Others have used intra- and inter-specific diversity comparisons to infer the introduction of transposable elements into host genomes (SANCHEZ-GRACIA et al. 2005; SILVA and KIDWELL 2000) and the diversity data for *Herves* is qualitatively similar to those data. Second, when elements are horizontally transferred to a new host species there is a period of time when natural selection will favor active autonomous elements and this will leave a distinct molecular signature within the elements in the form of a skewed ratio of synonymous and non-synonymous substitution rates (ROBERTSON and LAMPE 1995). In this study a comparison of the synonymous and non-synonymous substitution rates within the *Herves* transposase-coding region detected evidence of purifying selection and is consistent with the hypothesis that *Herves* was laterally introduced into this lineage from an unknown source.

Although *Herves* displays evidence of being horizontally introduced into the *An. gambiae* lineage, the timing of this event remains uncertain. The intensity of the molecular signals indicating horizontal transfer suggests that this event was not in the very recent past. Sanchez-Garcia et al (2005) recently examined 14 transposable elements in *D. melanogaster* and, based on silent site diversity, concluded that 13 were products of horizontal transfer that probably occurred approximately 5-12 million years ago. Sanchez-Garcia et al. (2005) observed levels of silent diversity within the transposable elements studied approximately 100-fold less than that observed in 21 nuclear genes while in this study silent site diversity was only 6-fold less than expected when the data were pooled, and ranged from 3-fold to 125-fold less than expected.
depending on the location from which the samples were collected. These data appear consistent with an historical lateral transfer event, although not one that has occurred recently.

The form diversity observed in this study is also consistent with Herves having an extended residence time within the An. gambiae lineage. Interestingly however, while the number of forms of Herves as determined by the sequence of the 5’ end of the transposase gene totaled 33, the frequency of individuals with at least one copy of an element that had either no internal deletions or deletions less than 100 bp (the limits of the detection method) was over 90%. Internally deleted elements can arise quickly following the introduction of a transposable element as has been displayed by the well-studied P element in Drosophila species (O’HARE et al. 1992). This is distinctly not the case for Herves and may be due to a number of factors. First, if deleted elements are preferentially removed from the genome then one would see a relative abundance of intact forms as observed here. Currently there are no data for the differential removal of smaller, internally deleted forms of an element and indeed, smaller non-autonomous elements can have an activity advantage in the presence of functional transposase (LAMPE et al. 1998; SPRADLING 1986). An alternative possibility is that Herves elements may have reduced opportunities to form internally deleted elements. Internal deletions of Class II transposable elements arise in some cases during the double-stranded DNA gap repair process following element excision. For example, following P element excision in D. melanogaster the resulting double-stranded gap is filled during a homology-dependent recombination process in which homologous or ectopic copies of a P element are copied into the gap (ENGELS et al. 1990). Premature resolution of these recombination products before this templated gap repair process is complete results in the creation of incomplete elements. The extent to which post-excision repair involves homology-dependent recombination or non-homologous end joining will determine, to
some extent, how often internally deleted elements are created within a genome (Rio 2002). A preference for Herves excision products to be repaired using non-homologous end joining mechanisms could explain two aspects of Herves observed in An. gambiae – the relative abundance of intact elements and their low copy number.

hAT element excision results in double-stranded breaks in the chromosome in which the ends of each chromosome are sealed by hairpin structures (Zhou et al. 2004). These hairpin structures are resolved by a nicking event followed by end-joining. The hairpin structures that arise on the empty donor site following hAT element excision are not seen following P element excision. We propose that this predisposes Herves post-excision repair to occur via non-homologous end-joining and thereby reduces the frequency with which internally deleted elements are created.

Herves is present at low copy numbers within An. gambiae and the data suggest that copy-number equilibrium has not been reached (Tajima’s D statistic for pooled data = -1.91). The low copy number of Herves, while not unique among Class II transposable elements, tends to be somewhat unexpected if the element was introduced into this lineage in the distant past. Class II transposable elements tend to increase in copy number when they are active within a genome. This increase in copy number occurs despite the conservative cut-and-paste nature of Class II element movement because the double-stranded breaks that arise following element excision can be repaired using homology dependent repair processes that result in a copy of the element being inserted into the gap (Rio 2002). Alternatively, an increase in copy number can occur as a result of Class II transposable elements moving from replicated regions of the genome to unreplicated regions of the genome during S-phase (Wilson et al. 2003). Although the mechanisms of copy number increase may vary, it seems well established that element copy-
number is expected to increase during periods of element activity. The low number of *Herves* elements in all individuals sampled therefore seems at odds with the diversity data that points to an extended residence time in the *An. gambiae* lineage. The tendency of different Class II transposable elements to increase in copy number has never been systematically compared although it is reasonable to think that some elements might be more “replicative” than others. *hAT* elements, and *Herves* in particular, may have a relatively low replication potential because of the presence of hairpin-containing intermediates following excision.

The structure of the population of *An. gambiae* in Africa has been studied and it has been proposed that there are two main divisions of the gene pool – a northwestern division including Senegal, Ghana, Nigeria, Cameroon, Gabon, Democratic Republic of Congo and western Kenya, and a southeastern division including Kenya, Tanzania, Malawi and Zambia (Lehmann *et al.* 2003). It has been proposed that there has been a recent bottleneck in the southeast division resulting in reduced genetic diversity followed by colonization from the northwest division. (Lehmann *et al.* 2003). The data presented here shows little evidence of geographical variation and the variation shown is inconsistent with the above model. Samples from Mozambique showed the highest levels of silent site diversity and no reduction in the diversity of forms as might be expected following a bottleneck. In fact, samples from Nigeria not only showed the least silent site diversity but also had the least amount of form diversity. Further sampling of *Herves* from populations in western Africa is needed to confirm the modest trends revealed in this study.

ACKNOWLEDGEMENTS

The generosity of Tovi Lehmann, Derek Charlwood, Christopher Curtis and Wilhelmine Meeraus for providing samples is gratefully acknowledged. Matthew Hare, Sky Lesnick, Floyd
Reed and Subhamoy Pal provided us with valuable advice and useful discussions. The National Institutes of Health, R01GM48102, supported this work.
## TABLE 1

**Site occupancy**

<table>
<thead>
<tr>
<th>Location</th>
<th>Na</th>
<th>Nb</th>
<th>dcn</th>
<th>βd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>24</td>
<td>25</td>
<td>3.5</td>
<td>9.5</td>
</tr>
<tr>
<td>Kisian</td>
<td>15</td>
<td>14</td>
<td>2.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Malindi</td>
<td>25</td>
<td>17</td>
<td>3.4</td>
<td>11.0</td>
</tr>
<tr>
<td>Zenet</td>
<td>73</td>
<td>31</td>
<td>3.8</td>
<td>2.1</td>
</tr>
<tr>
<td>Furvela</td>
<td>49</td>
<td>23</td>
<td>4.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>32</td>
<td>20</td>
<td>3.3</td>
<td>2.3</td>
</tr>
</tbody>
</table>

a Individuals analyzed by transposable element display  

b Number of unique chromosomal sites containing *Herves*  

c Diploid copy number of *Herves* (WRIGHT et al. 2001).  

d $4N_e(v+s)$ from Charlesworth and Charlesworth (1983).  

e Data from O’Brochta et al. (2006)
TABLE 2

Nucleotide sequence polymorphism in *Hervis*

<table>
<thead>
<tr>
<th>Location</th>
<th>Seqs&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Poly&lt;sup&gt;b&lt;/sup&gt;</th>
<th>π&lt;sup&gt;c&lt;/sup&gt;</th>
<th>θ&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Poly&lt;sup&gt;b&lt;/sup&gt;</th>
<th>π&lt;sup&gt;c&lt;/sup&gt;</th>
<th>θ&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>51</td>
<td>44</td>
<td>0.0056 (0.0037)</td>
<td>0.0076 (0.0023)</td>
<td>15 (3+12)</td>
<td>0.0034 (0.0042)</td>
<td>0.0063 (0.0023)</td>
</tr>
<tr>
<td>Kisian</td>
<td>29</td>
<td>60</td>
<td>0.0086 (0.0009)</td>
<td>0.0128 (0.0043)</td>
<td>7 (1+6)</td>
<td>0.0024 (0.0004)</td>
<td>0.0034 (0.0016)</td>
</tr>
<tr>
<td>Malindi</td>
<td>28</td>
<td>44</td>
<td>0.0076 (0.0006)</td>
<td>0.0084 (0.0029)</td>
<td>7 (2+5)</td>
<td>0.0033 (0.0005)</td>
<td>0.0034 (0.0016)</td>
</tr>
<tr>
<td>Zenet</td>
<td>57</td>
<td>109</td>
<td>0.0084 (0.0008)</td>
<td>0.0177 (0.0050)</td>
<td>21 (7+14)</td>
<td>0.0057 (0.0009)</td>
<td>0.0104 (0.0035)</td>
</tr>
<tr>
<td>Furvela</td>
<td>33</td>
<td>35</td>
<td>0.0091 (0.0004)</td>
<td>0.0079 (0.0027)</td>
<td>8 (5+3)</td>
<td>0.0056 (0.0032)</td>
<td>0.0037 (0.0017)</td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>40</td>
<td>53</td>
<td>0.0086 (0.0006)</td>
<td>0.0095 (0.0030)</td>
<td>6 (1+5)</td>
<td>0.0015 (0.0003)</td>
<td>0.0028 (0.0014)</td>
</tr>
<tr>
<td>Combined</td>
<td>238</td>
<td>124</td>
<td>0.0079 (0.0003)</td>
<td>0.0216 (0.0049)</td>
<td>35 (14+21)</td>
<td>0.0046 (0.0004)</td>
<td>0.0134 (0.0035)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of sequences analyzed

<sup>b</sup> Number of polymorphic positions; Numbers in parenthesis = synonymous + non-synonymous sites

<sup>c</sup> Pairwise nucleotide diversity (Nei and Li 1979); standard deviation in parenthesis

<sup>d</sup> Nucleotide diversity based on segregating sites (Waterson 1975); standard deviation in parenthesis
### TABLE 3

Genetic diversity of *Herves* elements from different locations

<table>
<thead>
<tr>
<th>Locations</th>
<th>Haploid copy number</th>
<th>(\pi_e^a)</th>
<th>Observed</th>
<th>Expected</th>
<th>(\pi_e^b)</th>
<th>Observed/Expected</th>
<th>Tajima(\hat{D})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>1.8</td>
<td>0.002</td>
<td>0.038</td>
<td>0.053</td>
<td>-1.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kisian</td>
<td>1.55</td>
<td>0.001</td>
<td>0.032</td>
<td>0.031</td>
<td>-0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malindi</td>
<td>1.7</td>
<td>0.002</td>
<td>0.036</td>
<td>0.056</td>
<td>-1.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zenet</td>
<td>1.9</td>
<td>0.005</td>
<td>0.040</td>
<td>0.126</td>
<td>-1.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furvela</td>
<td>2.15</td>
<td>0.015</td>
<td>0.045</td>
<td>0.334</td>
<td>1.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>1.7</td>
<td>0.0003</td>
<td>0.036</td>
<td>0.008</td>
<td>-1.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>1.8^c</td>
<td>0.006</td>
<td>0.038</td>
<td>0.158</td>
<td>-1.91*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^a\(\pi_e\) represents the average pairwise nucleotide diversity at synonymous sites.

^b see Material and Methods.

^c Average haploid copy number from all locations

^d \(P > 0.05\)

* \(P < 0.05\)
## TABLE 4

### Frequency of Open Reading Frames

<table>
<thead>
<tr>
<th>Location</th>
<th>N&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Complete ORF&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>24</td>
<td>1.00</td>
</tr>
<tr>
<td>Kisian</td>
<td>15</td>
<td>0.90</td>
</tr>
<tr>
<td>Malindi</td>
<td>25</td>
<td>0.88</td>
</tr>
<tr>
<td>Zenet</td>
<td>73</td>
<td>0.84</td>
</tr>
<tr>
<td>Furvela</td>
<td>49</td>
<td>0.95</td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>32</td>
<td>0.44</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of mosquitoes analyzed

<sup>b</sup> Frequency of mosquitoes with evidence of an intact *Hervas* ORF (2.1 kb PCR product).
TABLE 5

*Herves* ORF Form diversity

<table>
<thead>
<tr>
<th>Location</th>
<th>Seqs&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Forms</th>
<th>Form diversity&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>51</td>
<td>12</td>
<td>0.857 (0.028)</td>
</tr>
<tr>
<td>Kisian</td>
<td>29</td>
<td>9</td>
<td>0.820 (0.055)</td>
</tr>
<tr>
<td>Malindi</td>
<td>28</td>
<td>8</td>
<td>0.841 (0.044)</td>
</tr>
<tr>
<td>Zenet</td>
<td>57</td>
<td>17</td>
<td>0.903 (0.022)</td>
</tr>
<tr>
<td>Furvela</td>
<td>33</td>
<td>5</td>
<td>0.706 (0.049)</td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>40</td>
<td>7</td>
<td>0.565 (0.088)</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td><strong>238</strong></td>
<td><strong>33</strong></td>
<td><strong>0.833 (0.018)</strong></td>
</tr>
</tbody>
</table>

<sup>a</sup> Sequences analyzed

<sup>b</sup> Standard deviation in parenthesis
TABLE 6

Shared Forms between locations\textsuperscript{a}

<table>
<thead>
<tr>
<th></th>
<th>Asembo</th>
<th>Kisian</th>
<th>Malindi</th>
<th>Zenet</th>
<th>Furvela</th>
<th>Bakin-Kogi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asembo</td>
<td>5\textsuperscript{b}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kisian</td>
<td>4</td>
<td>4\textsuperscript{b}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malindi</td>
<td>5</td>
<td>4 \textsuperscript{b}</td>
<td>1 \textsuperscript{b}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zenet</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>10\textsuperscript{b}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furvela</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2 \textsuperscript{b}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bakin-Kogi</td>
<td>3</td>
<td>4 \textsuperscript{b}</td>
<td>4</td>
<td>5</td>
<td>2 \textsuperscript{b}</td>
<td>2 \textsuperscript{b}</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Number of Forms shared between locations

\textsuperscript{b} Number of Forms found at only this location
FIGURE LEGENDS

Figure 1. Political map of Africa showing locations of sample populations.

Figure 2. Site occupancy frequency distribution. A-F. The number of sites that were found in a sample exactly “x” times is plotted on the x-axis and the site occupancy is plotted on the y-axis.

Figure 3. Nucleotide polymorphism in Herves. The results of a sliding window analysis (100 bp window in 25 bp steps) showing the levels of nucleotide polymorphism, π, as a function of position within the element. The horizontal dotted line represents the average nucleotide polymorphism reported for 35 An. gambiae nuclear genes (MORLAIS et al. 2004). ITR, inverted terminal repeat; I, II, subterminal direct repeats; ORF, transposase open reading frame.

Figure 4. Network of genealogical relationships of forms of Herves ORFs based on statistical parsimony (TEMPLETON et al. 1992). The abundance and relationship of individual forms are shown. Each node represents a single mutational step. The area of the circles is proportional to the form frequency class. Shading refers to the region in which forms were found. In cases where forms are shared among regions, shading is proportional to the frequency of the form in each region. Small black dots represent missing forms.

Figure 5. Frequency of classes of Herves forms. Herves forms were classified based on the number of locations at which they were found (1-6). The number of forms in each class is plotted on the x-axis.
Figure 1.
Figure 2.
Figure 3.
Figure 4.
Figure 5.
LITERATURE CITED


ANXOLABEHERE, D., M. G. KIDWELL and G. PERIQUET, 1988 Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements. Mol Biol Evol **5**: 252-269.


Supplementary Figure 1:
Neighbor Joining (NJ) tree of the thirty three different forms of *Herves* based on the first 528 bp of the 5' end of the transposase open reading frame.