WR 371

Short communication

Intron retention identifies a malaria vector within the *Anopheles* (*Nyssorhynchus*) *albitaris* complex (Diptera: Culicidae) ☆

T.J.S. Merritt [*,1], C.R. Young [1,2], R.G. Vogt, R.C. Wilkerson [3], J.M. Quattro

Department of Biological Sciences, Program in Marine Science, Baruch Institute and School of the Environment,
University of South Carolina, Columbia, SC 29208, USA

1. Introduction

Taxonomic identification of cryptic mosquito species is more than an academic exercise when members of the species complex function as human disease vectors. The correct identification of species involved in disease transmission is vital for development of public health strategies and assessment of effectiveness of previously implemented strategies (e.g., vaccination or vector control efforts). Primary disease vectors may shift in response to variation in local environments including changes driven by human activity. For example, *Anopheles* (*Nyssorhynchus*) *marajoara* Galvão and Damasceno (Linthicum, 1988) is the principal malaria vector in northeastern Amazonia, replacing *An. darling* Root, perhaps as a result of changes in human activity (Conn et al., 2002). Discrimination of *An. marajoara* from other members of the cryptic Albitarsis Complex remains a challenge, and evolutionary relationships among the members of this complex remain unresolved.

Anopheles marajoara is one of four species in the Neotropical Albitarsis Complex which includes *An. albitarsis* Lynch Arribalzaga, *An. deaneorum* Rosa-Freitas, *An. marajoara*, and *An. albitarsis*. Morphological identification of adult females of the Albitarsis species complex is possible. However, species within this group can only be reliably distinguished using random amplified polymorphic DNA-polymerase chain reaction techniques (RAPD-PCR, Wilkerson et al., 1995a,b). An accurate and reliable method of identifying *An. marajoara* is needed due to the new prominence of this species as a mode of disease transmission.

In this study, we examined phylogenetic relationships within the Albitarsis Complex using a region of the *white* gene. This region contains both coding (exon) and noncoding (intron) sequence and has been identified as a promising gene region for the phylogenetic resolution of mosquito species (Besansky and Fahey, 1997). Four introns were originally described in mosquitoes (Besansky and Fahey, 1997). More recently, this number was found to vary among mosquito species, including the loss of the fourth intron in at least one member of the *An. albitarsis* species group, *An. albitarsis* (Krywinski and Besansky, 2002; Krywinski et al., 2001).

We report that within this species group, only *An. marajoara* retains the fourth intron of the *white* gene. We propose credible sets of species phylogenies for the members of the *An. albitarsis* complex and use flanking coding sequence to determine the statistical support for topologies consistent with a single intron loss during the evolutionary history of the *An. albitarsis* species complex. The presence/absence of this intron can be used to quickly and easily distinguish *An. marajoara*, the principal malaria vector in northeastern Amazonia, from other members of the *An. albitarsis* complex.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAR 2005**	2. REPORT TYPE	3. DATES COVERED **00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Intron retention identifies a malaria vector within the Anopheles (Nyssorhynchus) albitaris complex (Diptera: Culicidae)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Walter Reed Army Institute of Research,Department of Entomology,Silver Spring,MD,20910**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

2. Materials and methods

2.1. Samples

Individual female mosquitoes were collected and progeny broods reared as described in Wilkerson et al. (1995a,b). Individual families were identified within the Albitarsis Complex using RAPD profiles of emerging adults (Wilkerson et al., 1995a,b). DNA samples were taken from a minimum of two lines from each species. *An. albitarsis* occurs from southern Brazil to Buenos Aires, and we include specimens collected from Argentina, Misiones Province, Posadas (AR 1(7)) and from Brazil, Santa Catarina State, Macaranduba (BR 010(1)) in our phylogenetic analyses. *An. deaneorum* were collected from a site in the southern part of its range in Argentina, Corrientes Province, Corrientes (AR 3(4)), and from the type locality in Brazil, Rondônia State, Guajara-Mirim (BR/R 007(2)). *An. albitarsis* species B specimens were collected from the extremes of its distribution. One line was collected from Paraguay, Alto Paraná State, Hernandarias (PA 2(3)) and one line from Brazil, Pará State, Primavera (BR 009(1)). *An. marajoara* posseses a widespread distribution. We collected two lines from widely separated locations in Brazil, São Paulo State, Iguape (BR 73(14)), and Rondônia State, Costa Marques (BR 525(11)). A map showing all collections sites can be found in Wilkerson et al. (1995a).

2.2. White locus amplification and sequencing

Total DNA was purified from adult individuals using a commercially available kit (DNeasy, Qiagen). The region of the *white* locus flanking the fourth intron (as numbered in Besansky and Fahey, 1997) was amplified using a previously reported set of oligonucleotide primers (WZ2E and WZ11X Besansky and Fahey, 1997). PCR was carried out for 40 cycles under the following conditions: denaturation at 95 °C for 1 min, annealing at 48 °C for 1 min, and extension at 72 °C for 1 min. PCR products were sequenced using a 377 Automated Sequencer (Perkin-Elmer). Dye terminator sequencing was performed with BigDye termination mix (Perkin-Elmer Biosystems) using the manufacturer's suggested protocol. Two to fifteen samples per species were sequenced on both strands.

2.3. Phylogenetic analyses

In addition to the eight mosquito *white* sequences reported here, we obtained the *Anopheles albimanus white* sequence (U73839) from GenBank and used it to root phylogenies. We aligned coding-region nucleotides based on an alignment of inferred amino acid sequences produced using the ClustalW multiple alignment program (Thompson et al., 1994) and aligned the two *An. marajoara* intron sequences by eye. The resulting alignment consisted of 735 bp of the coding region after removal of gaps and the intron. All subsequent phylogenetic analyses were conducted on this reduced data set. We determined percent G + C using DnaSP (Rozas et al., 2003) and used the method of Rzhetsky and Nei (1995) to test the assumption of homogeneity of base composition among taxa.

We tested a set of nested models of sequence evolution that are restrictions of the general time reversible (GTR) model (Tavaré, 1986; Yang, 1994) with six substitution classes, unequal base frequencies, and rate variation among sites. The GTR model and its nested sub-models assume stationarity and a constant substitution process through time. To choose reasonable candidate models for the parametric Bayesian and likelihood analyses, we assumed a topology (neighbor-joining) and maximized the likelihood function for the candidate models using standard procedures in PAUP (ver. 4.0b10, Swofford, 1998). The relative fit of models was assessed by the Akaike information criterion, $AIC = -2 \ln L + 2n$, where L is the maximum likelihood score and n is the number of free parameters of the model. Smaller values of AIC are preferred (Akaike, 1974; Posada and Crandall, 2001).

We used PAUP to calculate likelihood scores for various phylogenetic models, to estimate pairwise sequence divergences using the best-fit model, and to conduct phylogenetic analyses. We conducted phylogenetic analyses using the methods of maximum parsimony, neighbor-joining, and maximum likelihood (ML) employing exhaustive searches where applicable (Felsenstein, 1981). All methods produced the same topology but varied in statistical support. We report in detail only the results of the likelihood and Bayesian analyses. We compared maximum likelihood and alternative topologies under the best-fit model using the approximately unbiased (AU) test (Shimodaira and Hasegawa, 1999). We used PAUP to calculate site-likelihood scores under the best-fit model for the AU test. Three unrooted topologies and 15 rooted topologies are possible with four taxa. Only eight topologies were considered because some topological hypotheses produced branches of effectively zero length, yielding unresolved polytomies. In these cases, site-likelihood scores are identical for topological hypotheses consistent with a particular polytomy, and rejection of the polytomy was considered a rejection of all topological hypotheses consistent with that polytomy.

We used MrBayes v3.0 (Huelsenbeck and Ronquist, 2001) to determine posterior probabilities of topology bipartitions. The best-fit model of substitution was assumed. Site specific rate parameters (diffuse, Dirichlet(1,1,1) prior) were assigned to partitions of first, second and third codon positions. Topology (diffuse prior, all topologies equally weighted), branch lengths (diffuse, uniform(0,100) prior), instantaneous rate matrices (diffuse, Dirichlet(1,1,1,1,1,1) prior), and equilibrium base frequencies (diffuse, Dirichlet(1,1,1,1) prior) were shared among the partitions.

Markov chain Monte Carlo (MCMC) convergence was assessed by visually inspecting the sample paths of model parameters and by repeating the analysis multiple times to assess convergence. For each analysis, the chain was iterated a minimum of 5.1×10^7 times, sampling parameters every 1000 iterations for the MCMC data set to reduce autocorrelation of the samples. Of these 51,000 samples, 1000 were discarded to ensure that the chain had reached convergence before inference from the MCMC data set was made. All repetitions of the analysis converged on very similar parameter estimates. Phylogenies were ranked according the their posterior probabilities, $P(\tau|X,\theta)$, where τ is a topology, X is the observed data matrix and θ is the vector of parameters of the substitution model. The posterior probabilities were used to construct 95 and 99% credible sets of phylogenies.

3. Results

PCR amplification using the WZ2E and WZ11X oligonucleotide primers resulted in a single fragment in all samples. Direct sequencing confirmed that this fragment was from the region of the *white* gene flanking the fourth intron of the *white* locus (GenBank Accession Nos.: AY956295–AY956302). Fragment length varied among species and across the *An. marajoara* sequences. The largest difference in fragment size was found between the *An. marajoara* sequences and the sequences from the other three species in the *An. albitarsis* species complex (84 or 87 bases, depending on the *An. marajoara* sequence). Examination of the putative coding regions indicated that this difference in size is the result of the exact excision of the fourth intron from the *An. albitarsis*, *An. albitarsis* sp.B and *An. deaneorum* sequences. Examination of 13 additional *An. marajoara* lines from a total of four locations across Brazil (Mato Grosso State, Peixoto de Azevedo, Amapá State, Macapá and the two previously noted locations) found that the intron was present in all *An. marajoara* populations sampled. Small differences in the length of the fourth intron (≤ 4 bp) were observed among the 15 *An. marajoara* samples. Small differences (3–15 bp) in fragment size across species were the result of insertion/deletion events within

the coding regions of the gene and were confined to a previously described (Besansky and Fahey, 1997) highly variable region. No insertion/deletions result in a frame shift.

A total of 107 sites out of 735 bp were polymorphic. Fifteen of these sites occurred in the first codon position, seven in the second and 85 in the third. There were a total of 14 inferred amino acid substitutions, most of which occurred in the aforementioned highly variable region. Percent G + C was very similar to previous observations for *An. albimanus* and *An. albitarsis* (Krywinski et al., 2001), and showed the same degree of bias in third positions for high G + C content (86%: ingroup, 83%: outgroup). Percent G + C in first positions (53%: ingroup, 50%: outgroup) resembled that of the intron of *An. marajoara* (59%), and percent G + C in second positions was lower (41%: ingroup, 42%: outgroup). Among the taxa considered in our study, G + C content is homogeneous (range: 59–61%; see also Table 1). Homogeneity of nucleotide frequencies was not rejected ($I = 31.92$, $P = 0.129$).

The best-fit model of sequence evolution was the GTR model with site-specific rates (GTR + SS, AIC = 3473.8). Selection of this model was not sensitive to the topology used to compute the likelihood scores (data not shown). Parameter estimates for the best-fit model are presented in Table 1. The ML point estimates in Table 1 were used in all subsequent ML analyses. Pairwise sequence divergences within species of the *An.* Albitarsis Complex were low (average: 0.002, range: 0.000–0.004, using the GTR + SS model) compared to that between species (average: 0.061, range: 0.047–0.080). *An. albimanus* differs from members of the *An. albitarsis* complex by an average of 0.113 (range: 0.092–0.125). Fig. 1A presents the maximum likelihood tree ($-\ln L = 1624.33$) with maximum parsimony bootstrap percentages reported below branches. All phylogenetic inference methods produced the same topology (Fig. 1A).

Our analyses place *An. marajoara* as the basal taxon in the species group. Given the frequent gain and loss of introns of the *white* locus of Anophelinae (Krywinski et al., 2001), it is possible that the apparent loss in this group could in fact represent multiple, independent, loss events. It is desirable, therefore, to determine statistical

Table 1
Parameters for phylogenetic analyses

	Rate Matrix					Base Frequencies				SS rates		
	CT	CG	AT	AG	AC	p_A	p_C	p_G	p_T	m_1[a]	m_2[a]	m_3[a]
MLE	8.6	3.5	0.55	9.1	2.8	0.17	0.32	0.29	0.21	0.39	0.14	2.5
Median	10.5	4.6	0.26	11.4	4.2	0.18	0.33	0.29	0.21	0.41	0.20	2.4
Lower[b]	3.9	1.6	0.01	4.1	1.4	0.15	0.30	0.26	0.18	0.26	0.11	2.2
Upper[b]	50.9	22.8	4.23	53.6	20.4	0.20	0.36	0.32	0.23	0.60	0.34	2.6

[a] Indicator definitions: m_1: 1st-pos.; m_2: 2nd-pos.; m_3: 3rd-pos.

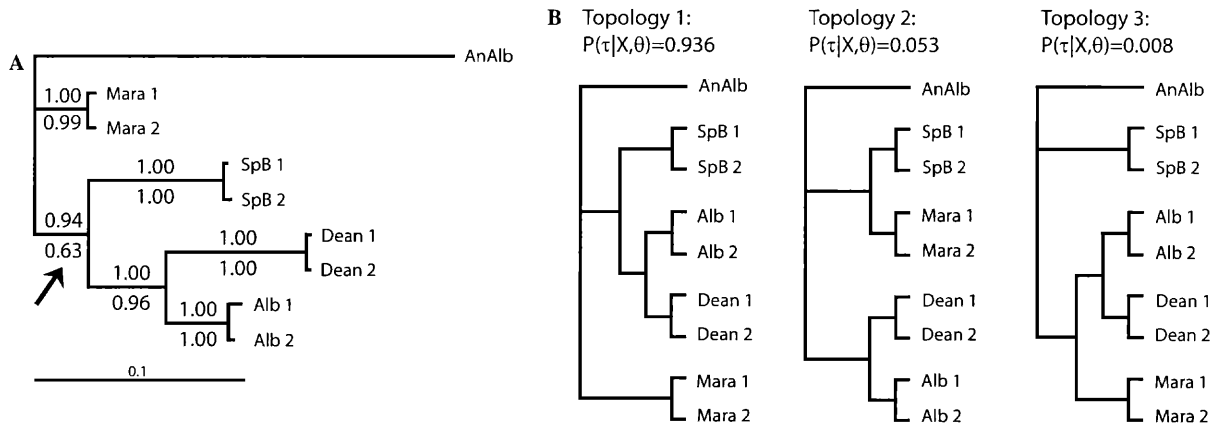[b] 95% credible intervals calculated from marginal posterior distributions.

Fig. 1. (A) Phylogenetic relationships within the *Anopheles albitarsis* species group based on *white* coding sequences. Taxon designations: AnAlb, *An. albimanus*; Mara, *An. marajoara*; sp.B, *An. albitarsis* species B; Dean, *An. deaneorum*; and Alb, *An. albitarsis*. The arrow indicates the phylogenetic location of the loss of the fourth intron assuming a single loss event. Numbers above and below each branch represent, respectively, the posterior probability of the taxa bipartition (GTR + SS) and parsimony bootstrap support. (B) Topologies included in the 95% (Topologies 1 and 2) and 99% (Topologies 1, 2, and 3) credible sets based on posterior probabilities conditional on the GTR + SS model. Species designations are identical to (A). The marginal posterior probabilities of each topology are given above each tree.

support for the placement of *An. marajoara* as basal to the other members of the Albitarsis Complex. Placement of any other species as the basal taxon would require that the fourth intron has been lost multiple times within this group, though placement of *An. marajoara* as the basal taxon does not exclude the possibility that the intron has been lost multiple times. Placement of the root was not well supported in the ML analysis (Fig. 2), suggesting that the data are consistent with *An. marajoara* as the basal taxon (T = −2.7, P = 0.874), a sister relationship between *An. marajoara* and *An. albitarsis* sp.B (T = 2.7, P = 0.334) or *An. albitarsis* sp.B as the basal taxon (T = 3.3, P = 0.088). Strong support exists for a sister relationship of *An. albitarsis* and *An. deaneorum* (Fig. 2).

The Bayesian analysis, on the other hand, suggests that the data support a single topology (Tree 1 in Fig. 1B) with a posterior probability of 0.936. The odds that this topology is the true topology for this locus, conditional on the model and the data, is about 15:1. A 95% credible set of topologies contains two topologies (Topology 1 and Topology 2 in Fig. 1B), and a 99% credible set of topologies contains only three topologies (Topology 1, Topology 2, and Topology 3 in Fig. 1B). Topology 2 places *An. marajoara* as the sister species of *An. albitarsis* sp.B ($P(\tau|X, \theta) = 0.053$), whereas Topology 3 places *An. albitarsis sp. B* as the basal taxa of the Albitarsis Complex ($P(\tau|X, \theta) = 0.008$). Posterior probabilities of taxon bipartitions that cluster members of each species are very high ($P = 1.00$: all bipartitions) strongly supporting the species assignments previously based on RAPD analysis. The posterior probability that *An. marajoara* is basal to the other three species provides weak support for the topology consistent with a single intron loss event in the evolutionary history of the *Albitarsis* Complex.

4. Discussion

Our phylogenetic analyses are consistent with previous suggestions, based on RAPD profiles (Wilkerson et al., 1995a,b), defining four phylogenetically discrete taxa within the *An. albitarsis* complex and provide a simple method for identifying one species, *An. marajoara*, which has recently emerged as a dominant malaria vector (Conn et al., 2002). The pronounced difference in size of *white* locus fragments amplified from *An. marajoara* and those amplified from any other member of the *An. albitarsis* complex can be used as a simple, reliable method for discrimination of *An. marajoara*; PCR fragments electrophoretically separated on an ethidium bromide-stained gel can be used to distinguish *An. marajoara* from the other members of the *An. albitarsis* complex without direct sequencing or RAPD-PCR.

Until recently most studies lumped these four species into *An. albitarsis* or the *Albitarsis* Complex, thus obscuring which species were truly involved in malaria transmission. Vector status is definitively known for only *An. marajoara*. *An. deaneaorum* is assumed to be a malaria vector, and vector status is not known for *An. albitarsis* or *An. albitarsis* sp.B. The two putative most important vectors, *An. deaneaorum* and *An. marajoara* can now be separated by means other than RAPDs; *An. deaneorum* by a larval character and *An. marajoara* by PCR of the white gene. This assay is a considerable improvement over methods previously available to identify *An. marajoara*: adult females are not required; minimal time and money are required to screen large numbers of individuals; targeted PCR does not suffer from the difficulties encountered in RAPD-PCR, and minimal technical skill is required perform and interpret the assay.
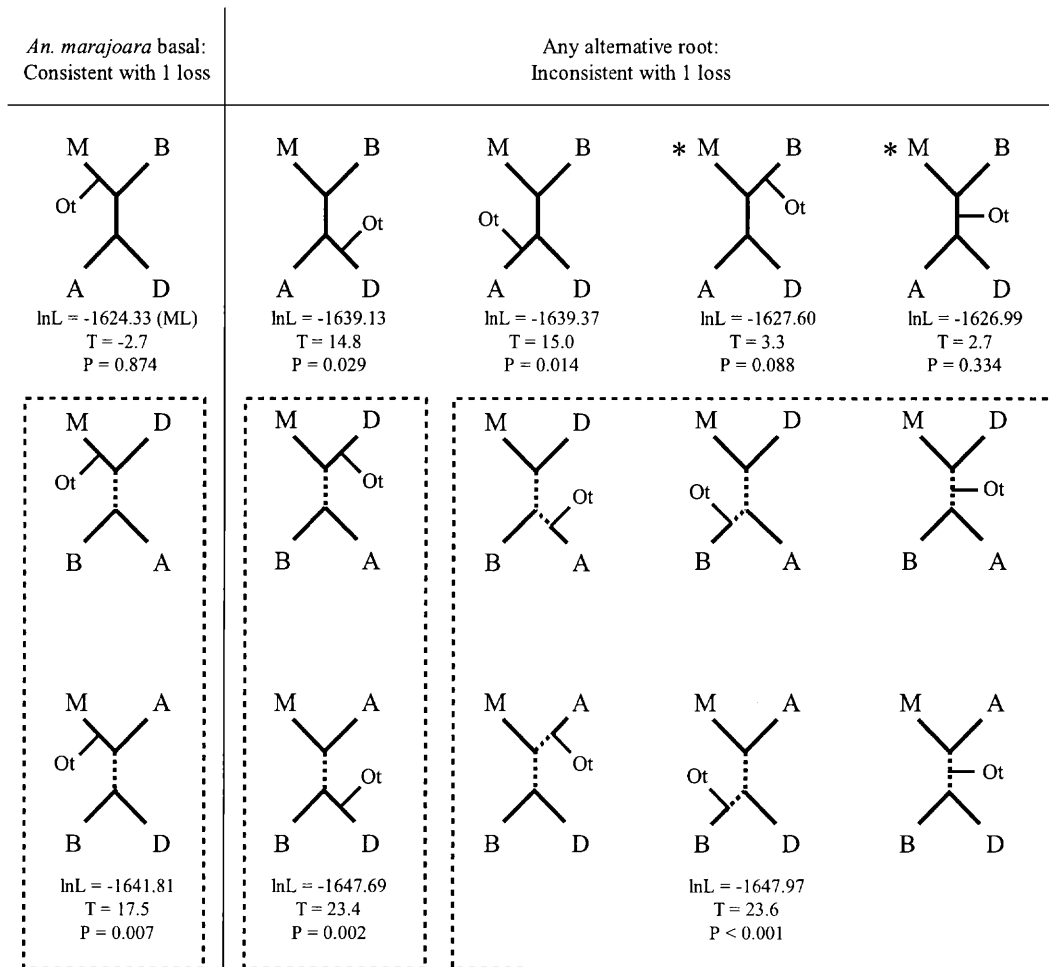
Fig. 2. AU tests of topologies. M, *An. marajoara*; A, *An. albitarsis*; B, *An. albitarsis* sp.B; D, *An. deaneorum*, and Ot, *An. albimanus*. Boxes indicate topologies that yield the same likelihood scores due to zero length branches (dashed branches). The left column contains topologies consistent with one loss of the *white* intron. All other topologies require multiple losses/gains. Asterisks indicate topologies that are not statistically distinguishable.

We propose a 95% credible set of species phylogenies for the *An. albitarsis* complex containing two topologies. Both topologies define a sister species relationship between *An. albitarsis* and *An. deaneorum*. We find weak evidence that *An. marajoara* is the basal taxon of this group, with the alternative topology placing *An. albitarsis* sp.B and *An. marajoara* as sister taxa. The topology with the highest posterior probability is consistent with a single loss of the fourth intron in this complex, whereas the alternative topology included in the 95% credible set is inconsistent with a single loss of the intron. Additional loci should be examined to test the phylogenetic hypotheses supported by this locus.

Acknowledgments

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contrib. 19, 716–723.

Besansky, N.J., Fahey, G.T., 1997. Utility of the white gene in estimating phylogenetic relationships among mosquitoes (Diptera: Culicidae). Mol. Biol. Evol. 14, 442–454.

Conn, J.E., Wilkerson, R.C., Segura, M.N., De Souza, R.T., Schlichting, C.D., Wirtz, R.A., Povoa, M.N., 2002. Emergence of a new neotropical malaria vector facilitated by human migration and changes in land use. Am. J. Trop. Med. Hyg. 66, 18–22.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Huelsenbeck, J.P., Ronquist, F., 2001. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8), 754–755.

Krywinski, J., Besansky, N.J., 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. Mol. Biol. Evol. 19, 362–366.

Krywinski, J., Wilkerson, R.C., Besansky, N.J., 2001. Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: insights from nuclear single-copy genes and the weight of evidence. Syst. Biol. 50, 540–556.

Linthicum, K.J., 1988. A revision of the Argyritarsis section of the subgenus *Nyssorhynchus* of *Anopheles*. Mosq. Syst. 20, 98–271.

Posada, D., Crandall, K.A., 2001. Selecting the best fit model of nucleotide substitution. Syst. Biol. 50, 580–601.

Rozas, J., Sanchez-Delbarrio, J.C., Messeguer, X., Rozas, R., 2003. Dnasp, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496–2497.

Rzhetsky, A., Nei, M., 1995. Tests of the applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12, 131–151.

Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16, 1114–1116.

Swofford, D.L., 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Sunderland, MA, Sinauer.

Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.

Wilkerson, R.C., Gaffigan, T.V., Lima, J.B., 1995a. Identification of species related to *Anopheles* (*Nyssorhynchus*) *albitarsis* by random amplified polymorphic DNA-polymerase chain reaction (Diptera: Culicidae). Mem. Inst. Osw. Cruz 90, 721–732.

Wilkerson, R.C., Parsons, T.J., Klein, T.A., Gaffigan, T.V., Bergo, E., Consolim, J., 1995b. Diagnosis by random amplified polymorphic DNA polymerase chain reaction of four cryptic species related to *Anopheles* (*Nyssorhynchus*) *albitarsis* (Diptera: Culicidae) from Paraguay, Argentina, and Brazil. J. Med. Entomol 32, 697–704.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39, 306–314.