Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 6885–6888, July 1994
Evolution

Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors

NORA J. BESANSKY*†, JEFFREY R. POWELL‡, ADALGISA CACCONE‡§, DIANE MILLS HAMM*, JULIE A. SCOTT*¶, AND FRANK H. COLLINS*†‖

*Division of Parasitic Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333; †Department of Biology, Emory University, Atlanta, GA 30322; ‡Department of Biology, Yale University, New Haven, CT 06511; and §Dipartimento di Biologia, II Universitá di Roma "Tor Vergata," 00173 Rome, Italy

Communicated by Fotis C. Kafatos, April 15, 1994

ABSTRACT **The six Afrotropical species of mosquitoes comprising the *Anopheles gambiae* complex include the most efficient vectors of malaria in the world as well as a nonvector species. The accepted interpretation of evolutionary relationships among these species is based on chromosomal inversions and suggests that the two principal vectors, *A. gambiae* and *Anopheles arabiensis*, are on distant branches of the phylogenetic tree. However, DNA sequence data indicate that these two species are sister taxa and suggest gene flow between them. These results have important implications for malaria control strategies involving the replacement of vector with nonvector populations.**

The *Anopheles gambiae* complex of mosquitoes comprises six species, all but one of which are involved in the transmission of human malaria parasites. Because of the marked anthropophily of *A. gambiae* and *Anopheles arabiensis*, this vectorial system is the most stable and deadly in the world. Long-range malaria control strategies are partly based on genetic modification of the capacity of the natural vector populations to transmit malaria parasites (1). Thus, establishing levels of reproductive isolation and phylogenetic relatedness of the members of the *A. gambiae* complex has important implications for the control of malaria.

The involvement of mosquitoes in malaria transmission is dependent on behavioral attributes, such as finding and biting of hosts and choice of resting and oviposition sites, that vary both within and between species (2). Correlated with behavioral differences are clinal geographic and microspatial variation in the frequencies of specific paracentric chromosomal inversions in polymorphic species (2–4), suggesting that alternative inverted arrangements have adaptive significance that may influence some of these behaviors. All but one member of the complex are polymorphic for paracentric inversions, but at least one fixed inversion differentiates pairs of sibling species (2). Indeed, chromosomal inversions may play an active role in Anopheline speciation (5) and form the basis for the accepted hypothesis concerning phylogenetic relationships in the *A. gambiae* complex (2).

The accepted phylogeny, which places the two principal vectors, *A. gambiae* and *A. arabiensis*, on distant branches of the phylogenetic tree, was inferred by parsimony analysis of paracentric inversions identified in polytene chromosomes (refs. 2 and 6; J.R.P. and A.C., unpublished data; see Fig. 1). Surprisingly, the species relationships predicted by this phylogeny contradict morphological, behavioral, ecological, and interspecies hybridization data, all of which agree that the species pairs *A. gambiae–A. arabiensis* and *Anopheles melas–Anopheles merus* are the most closely related (2, 7, 8). This conflict was explained by invoking evolutionary convergence of morphologic and behavioral traits (2). However, an alternative explanation is that the chromosomal phylogeny is misleading, because (*i*) the assumption that the inversion breakpoints were unique may be invalid, because either the breakpoints of cytologically identical inversions are not precisely the same and therefore were not generated by the same event, or they are precisely the same but occurred more than once; (*ii*) paracentric inversions may have been passed between species through introgression; and/or (*iii*) ancestral populations may have been polymorphic for chromosomal inversions.

Inferring the evolutionary relationships in this species complex poses a challenge to all available phylogenetic techniques, since several lines of evidence suggest that the *A. gambiae* complex represents one of the most recently diverged groups of sibling species yet studied. Nei's genetic distances range only between 0.10 and 0.25 across species (9), and DNA·DNA hybridization studies of total single-copy DNA were unable to separate the species within the complex (N.J.B., A.C., and J.R.P., unpublished data). Indeed, it has been suggested that speciation in these mosquitoes, particularly the most anthropophilic ones, has been driven by their adaptation to the environment of humans (2, 9), whose population density in tropical Africa markedly increased with the introduction of agriculture in the past 10,000 years (10). Crossing experiments in the laboratory have shown that reproductive isolation is incomplete (7), and interspecies hybrids are occasionally found in nature [0.1–0.2% (11)].

To test the predictions of the chromosomal phylogeny, we obtained molecular sequence data from nuclear ribosomal DNA (rDNA), mitochondrial DNA (mtDNA), and an esterase gene of five available taxa in the *A. gambiae* complex and performed phylogenetic analyses using maximum parsimony (MP), neighbor joining (NJ), and maximum likelihood (ML). The Asian mosquito *Anopheles sundaicus*, a close relative also in the Pyretophorus series, was used as an outgroup in the mtDNA analysis.

MATERIALS AND METHODS

Sample Preparation. Mosquito strains used in this study were obtained from laboratory colonies maintained at the Centers for Disease Control. The cloning and sequence determination of the nuclear rDNA intergenic spacer sequences (GenBank data base accession nos. U10135–U10139) have been described (12). For the mtDNA and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: indels, insertion/deletions; ML, maximum likelihood; MP, maximum parsimony; NJ, neighbor joining; rDNA, ribosomal DNA; Ti, transition(s); Tv, transversion(s).

¶Present address: Department of Entomology, University of Arizona, Tucson, AZ 85721.

‖To whom reprint requests should be addressed at: Entomology Branch, Mail Stop F-22, Division of Parasitic Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333.

esterase sequences (GenBank data base accession nos. U10123–U10134 and U10140–U10147, respectively), genomic DNA was extracted from individual adult mosquitoes (13) and amplified by PCR using the GeneAmp kit (Perkin–Elmer/Cetus) and following the manufacturer's protocol. Amplification parameters were 94°C for 1 min, 37°C for 2 min, and 72°C for 3 min for 35 cycles. The primers were based on the nucleotide sequence of the ND4–ND5 mtDNA genes determined from *A. gambiae* G3 (ref. 14; GenBank data base accession no. L20934)—DMP3B (plus strand), positions 7251–7270; DMP4A (plus strand) and DMP3A (minus strand), positions 7680–7699; DMP4B (minus strand), positions 8680–8695—or from nontranslated sequence downstream of an esterase gene from *A. gambiae* G3 (J.A.S. and F.H.C., unpublished data)—EST5S (plus strand), CTGTCGACCCAGACTGACTAAGCACTTTG; EST3B (minus strand), CTGGATCCATCGTACAACACACGTGCCC. PCR products were gel purified and cloned into pBluescript SKII+ (Stratagene). Sequencing was performed on double-stranded templates using the Sequenase 2.0 kit (United States Biochemical).

Sequence Analysis. Sequence analysis included 1849 bp of the rDNA intergenic spacers, 1445 bp of the ND4–ND5 mtDNA genes, and 571 bp of an esterase gene. Nucleotide sequences were aligned by using the PILEUP program of the Genetics Computer Group (GCG) (15). MP analysis was performed with PAUP3.0 (16), using the "branch and bound" option for tree searches. The PHYLIP3.4 (17) programs SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE were used for NJ analysis; DNAML was used for ML analysis. The subrepeat structure of the rDNA intergenic spacer sequences was determined by summarizing dot plots (produced by COMPARE and DOTPLOT in GCG) from all 10 pairwise sequence comparisons.

RESULTS AND DISCUSSION

The results of the analyses for the rDNA and species-consensus mtDNA data sets are summarized in Table 1. Because multiple strains of a species were sequenced for the mtDNA but not the rDNA data, consensus mtDNA sequences were calculated for each species by using polymorphic designations for sites variable among strains. In the mtDNA data, observed transitions/transversions (Ti/Tv) varied from 45:1 for intraspecific comparisons, 15:1 between *A. gambiae* and *A. arabiensis*, 5.6:1 among the other siblings, and 0.6:1 to *A. sundaicus*. In the rDNA data, Ti/Tv varied only from 1:1 to 2:1 for all comparisons. No insertion/deletions (indels) were found for the mtDNA sequences, but fourteen 1- to 7-bp indels requiring gaps in alignment occurred in the rDNA sequences, as well as one large 159-bp indel shared by *A. gambiae* and *A. arabiensis*. When gaps were included in the analysis, they were weighted equally as one character change; if more than two sized gaps existed at a site, this was coded as multiple character states. In the combined mtDNA and rDNA analysis, no weights were used and gaps were excluded. All weightings of both data sets yielded one MP tree except when no weights were given to rDNA data. The association of *A. gambiae* and *A. arabiensis* has strong statistical support from all three types of analysis, regardless of the weighting scheme used for Ti and Tv or the inclusion of indels.

Fig. 1 compares the chromosomal inversion phylogeny to the phylogenetic trees obtained by MP analysis of individual and combined data sets from the mtDNA and rDNA sequences. Although these trees are unrooted, the mtDNA data for *A. sundaicus* indicate that the root would be on the outermost branch (see Fig. 3). The indicated branch lengths are for Ti/Tv weights of 1:6 (mtDNA) and 1:1.5 (rDNA, excluding gaps). Each molecular phylogeny decisively clustered *A. gambiae* with *A. arabiensis*. By contrast, the chromosomal inversion phylogeny paired *A. gambiae* with *A. merus*, based on a shared X chromosome inversion unique to these species. In this regard, it is important to note that the rDNA genes in members of the *A. gambiae* complex are X chromosome linked (19) but outside the breakpoints of this inversion. No molecular data set could resolve the *Anopheles quadriannulatus–A. melas–A. merus* trichotomy. In fact, each of the three possible species pairs is weakly supported by a different data set. However, the subrepeat structure at

Table 1. Summary of phylogenetic analysis

	Ti/Tv weights								
				rDNA					
	mtDNA			Gaps			No gaps		
	None	1:3	1:6	None	1:1.5	Tv only	None	1:1.5	Tv only
	MP trees								
Tree length	65	75	90	248	292	122	231	283.5	105
No. of trees	1	1	1	2	1	1	2	1	1
Additional steps to									
mtDNA tree	—	—	—	+3	+4	+2	+3	+4	+2
rDNA tree	+1	+1	+1	—	—	—	—	—	—
Inversion tree	+8	+12	+18	+14	+17.5	+9	+12	+16.5	+7

	MP*			NJ		ML†	
	mtDNA	rDNA	mtDNA+rDNA	mtDNA	rDNA	mtDNA	rDNA
	Bootstrap values						
GAM–ARA	99–100	98–100	100	99	94	+	+
MEL–MER	11–14	13–54	18–51	27	62	–	–
MEL–QUA	62–72	2–19	11–17	50	3	–	–
MER–QUA	11–14	42–70	34–65	22	35	–	–

GAM, *A. gambiae*; ARA, *A. arabiensis*; MEL, *A. melas*; MER, *A. merus*; QUA, *A. quadriannulatus*.

*Ranges of values correspond to weighting schemes shown above.

†ML analysis was based on both Kimura's and Jukes and Cantor's estimates of multiple hits; Kimura estimates for mtDNA data included Ti/Tv weightings of 1:1, 1:3, 1:10; for rDNA, 1:1, 1:2. +, Breaking the association resulted in a significantly worse fit to the data according to Kishino and Hasegawa's test (18); neither the algorithm to estimate multiple hits nor weighting affected the significance of the test.

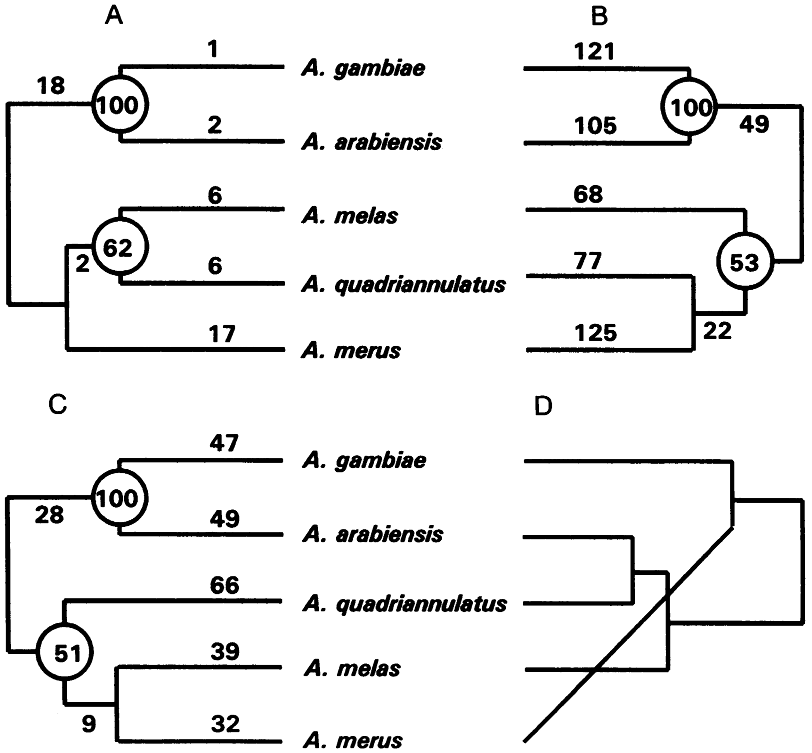

FIG. 1. MP trees based on mtDNA (*A*), rDNA (*B*), combined mtDNA and rDNA data sets (*C*), and chromosomal inversions (*D*) assuming a monophyletic origin of inversions. Numbers on branches are number of steps. Circled numbers are bootstrap values > 50 based on 100 replicates.

the 5′ end of the rDNA intergenic spacer (12), summarized schematically in Fig. 2, strongly corroborates the *A. melas–A. merus* and *A. gambiae–A. arabiensis* associations, with *A. quadriannulatus* showing affinity with the *A. melas–A. merus* clade.

When the mtDNA data are analyzed keeping individual strains separate, the MP tree does not support the monophyly of *A. gambiae* and *A. arabiensis*, but rather shows them intertwining (Fig. 3). Multiple strains of *A. gambiae* were also used to obtain sequences from a nuclear gene, the noncoding region 3′ to an esterase gene on chromosome 2L (J.A.S. and F.H.C., unpublished data). Unlike the mtDNA data, the esterase data have little resolving power among the sibling species, producing at least 10 most-parsimonious trees, one of which is shown (Fig. 3). Nevertheless, the esterase data do support monophyly of *A. gambiae* and *A. arabiensis*. This suggests that mtDNA but not esterase gene introgression has occurred and is consistent with the idea that mtDNA introgresses more readily than nuclear sequences (20–22).

While it is difficult to generate a robust phylogenetic hypothesis for species like the *A. gambiae* complex whose evolutionary histories have been so short, nuclear gene sequences from two different chromosomes and mtDNA gene sequences consistently and strongly support the relationship of *A. gambiae* and *A. arabiensis* as sister taxa. This is true regardless of the method of phylogenetic inference. We conclude that the chromosomal inversion phylogeny is not an accurate representation of species relationships in the *A. gambiae* complex. The discordance between the inversion- and gene sequence-derived species phylogenies may owe to some combination of introgression and polyphyletic or paraphyletic origin of inversions. *A. gambiae* and *A. arabiensis* share several chromosome 2 inversions, and two of these have been moved between species in laboratory experiments (23).

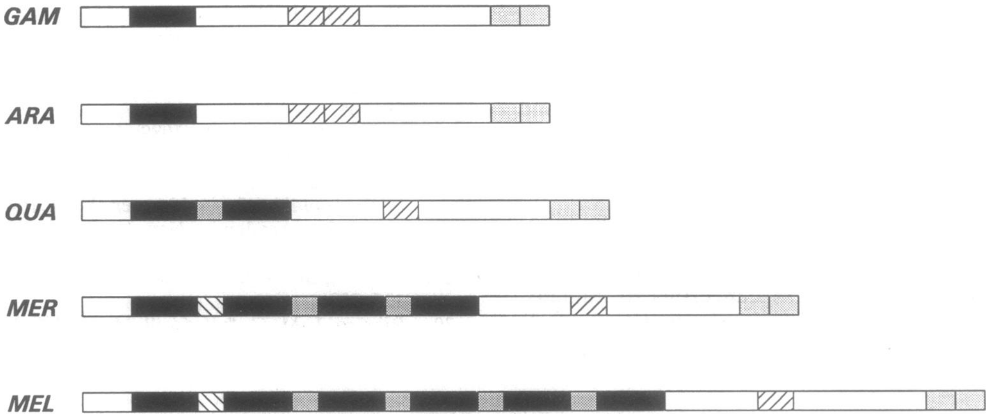

FIG. 2. Schematic representation of the subrepeat structure at the 5′ end of the rDNA intergenic spacer of five members of the *A. gambiae* complex. Abbreviations of species names are as in Table 1. Open boxes, unique sequence; solid, hatched, and stippled boxes, subrepeats. Identical fill patterns indicate related sequences.

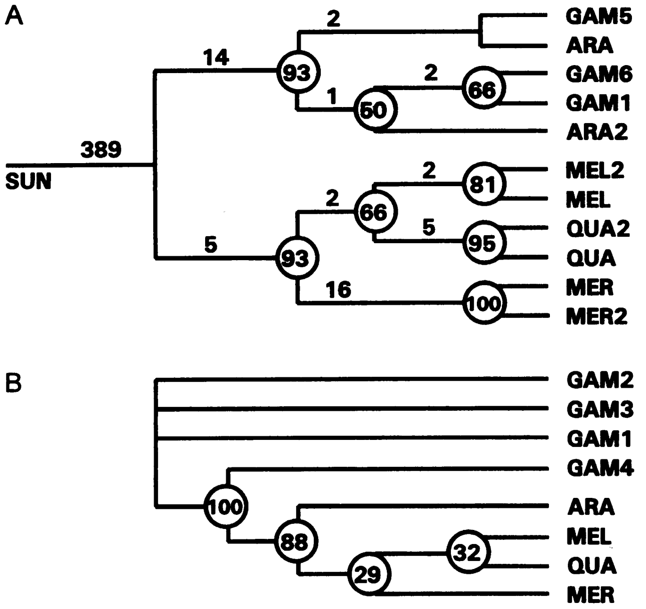

FIG. 3. MP trees inferred without generating a consensus from the sequences of individual strains. Abbreviations of species names are as in Table 1; SUN, *A. sundaicus*. Strains used were as follows—*A. gambiae*: GAM1, G3; GAM2, 4A; GAM3, L35; GAM4, SUA; GAM5, GMMK6; GAM6, MU; *A. arabiensis*: ARAB, ARZAG; ARAB2, GMAL; *A. melas*: MEL, BAL; MEL2, BRE; *A. merus*: MER, V12; MER2, ZULU; *A. quadriannulatus*: QUAD, SQUAD; QUAD2, CHIL. mtDNA (*A*) and esterase (*B*) trees used a Ti/Tv weighting of 1:6 and 1:1.5, respectively. Numbers on branches indicate branch lengths. Circled numbers are bootstrap values expressed as percentages based on 1000 replicates (mtDNA) or 100 replicates (esterase).

In addition, inversion breakpoints are not randomly distributed; ≈25% of the >30 described autosomal paracentric inversions in this species complex appear to share a breakpoint with another inversion, consistent with the presence of hot spots for double-strand breaks that could give rise to multiple occurrences of the same inversion. Moreover, selection is apparently maintaining inversion polymorphism in contemporary populations and may have done so in the past. It is therefore likely that inversion polymorphisms predated speciation events. If several lineages diverged from the same polymorphic ancestor, lineage sorting (24, 25) of inversions would make these rearrangements poor indicators of the phylogenetic history of the group.

Because many Diptera have excellent polytene chromosomes, chromosomal inversions have been benchmarks of phylogenetic inference within species groups (26, 27). However, the conflict between inversion phylogeny and species phylogeny is unlikely to be unique to the *A. gambiae* complex (28). A large number of Diptera that transmit human pathogens, especially other Anopheline vectors of malaria and blackfly vectors of filarial worms, are very closely related members of sibling species complexes for which no alternative genetic data exist. In addition to the inversion tree–species tree conflict, this study also underscores the importance of understanding phylogenetic relationships in designing control strategies that consider genetic manipulation of vector species to render them less dangerous to humans. Because a likely strategy for genetic manipulation of *A. gambiae* will be the use of an infectious engineered transposon or symbiont to drive antiparasite genes into populations (29, 30), even the limited genetic introgression observed between *A. gambiae* and *A. arabiensis* will impact the scope and dynamics of population replacement.

We thank Brian Holloway and the staff of the National Center for Infectious Diseases Biotechnology Core Facility for oligonucleotide synthesis, and Drs. Visut Baimai and Chaliow Kuvangkadilok for providing *A. sundaicus*. This study was supported by grants from the John D. and Catherine T. MacArthur Foundation, the United Nations Development Program/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases, and the National Centers for Infectious Diseases, Centers for Disease Control and Prevention.

1. World Health Organization (1991) *Prospects for Malaria Control by Genetic Manipulation of Its Vectors*, Report no. TDR/BCV/MAL-ENT/91.3 (World Health Organization, Geneva).
2. Coluzzi, M., Sabatini, A., Petrarca, V. & DiDeco, M. A. (1979) *Trans. R. Soc. Trop. Med. Hyg.* **73**, 483–497.
3. Coluzzi, M., Petrarca, V. & DiDeco, M. A. (1990) *Parassitologia Suppl.* **32**, 66–67.
4. Coluzzi, M. (1992) *Parasitol. Today* **8**, 113–118.
5. Coluzzi, M. (1982) in *Mechanisms of Speciation*, ed. Barigozzi, C. (Liss, New York), pp. 143–153.
6. Pape, T. (1992) *Mosq. Syst.* **24**, 1–11.
7. Davidson, G., Paterson, H. E., Coluzzi, M., Mason, G. F. & Micks, D. W. (1967) in *Genetics of Insect Vectors of Disease*, eds. Wright, J. W. & Pal, R. (Elsevier, Amsterdam), pp. 211–250.
8. White, G. B. (1974) *Trans. R. Soc. Trop. Med. Hyg.* **68**, 278–301.
9. Coluzzi, M., Petrarca, V. & DiDeco, M. A. (1985) *Boll. Zool.* **52**, 46–63.
10. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1993) *Science* **259**, 639–646.
11. White, G. B. (1971) *Nature (London)* **231**, 184–185.
12. Scott, J. A., Brogdon, W. G. & Collins, F. H. (1993) *Am. J. Trop. Med. Hyg.* **49**, 520–530.
13. Collins, F. H., Porter, C. H. & Cope, S. E. (1990) *Am. J. Trop. Med. Hyg.* **42**, 417–423.
14. Beard, C. B., Mills Hamm, D. & Collins, F. H. (1993) *Insect Mol. Biol.* **2**, 103–124.
15. Genetics Computer Group (1991) *Program Manual for the GCG Package* (Genetics Computer Group, Madison, WI), Version 7.
16. Swofford, D. (1991) PAUP: Phylogenetic Analysis Using Parsimony (Illinois Nat. Hist. Survey, Champaign, IL), Version 3.0.
17. Felsenstein, J. (1989) *Cladistics* **5**, 164–166.
18. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29**, 170–179.
19. Collins, F. H., Paskewitz, S. M. & Finnerty, V. (1989) *Adv. Dis. Vector Res.* **6**, 1–28.
20. Powell, J. R. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 492–495.
21. Ferris, S. D., Sage, R. D., Huang, C.-M., Nielsen, J. T., Ritte, U. & Wilson, A. C. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2290–2294.
22. Solignac, M., Monnerot, M. & Mounolou, J.-C. (1986) *J. Mol. Evol.* **23**, 31–40.
23. della Torre, A., Merzagora, L., Petrangeli, G. & Coluzzi, M. (1990) *Parassitologia* **32**, 80–81.
24. Avise, J. C., Neigel, J. E. & Arnold, J. (1984) *J. Mol. Evol.* **20**, 99–105.
25. Pamilo, P. & Nei, M. (1988) *Mol. Biol. Evol.* **5**, 568–583.
26. Dobzhansky, T. & Sturtevant, A. H. (1938) *Genetics* **23**, 28–64.
27. Carson, H. L. & Yoon, J. S. (1980) in *The Genetics and Biology of Drosophila*, eds. Ashburner, M., Carson, H. L. & Thompson, J. N. (Academic, New York), Vol. 3b, pp. 298–344.
28. Powell, J. R. (1991) *Mol. Biol. Evol.* **8**, 892–896.
29. Kidwell, M. G. & Ribeiro, J. M. C. (1992) *Parasitol. Today* **8**, 325–329.
30. Ribeiro, J. M. C. & Kidwell, M. G. (1994) *J. Med. Entomol.* **31**, 10–16.