npg

ORIGINAL ARTICLE

Mitochondrial DNA variation in the malaria vector *Anopheles minimus* across China, Thailand and Vietnam: evolutionary hypothesis, population structure and population history

B Chen[1,6], PM Pedro[2,6], RE Harbach[3], P Somboon[4], C Walton[2,6,7] and RK Butlin[5,6,7]

[1]*Institute of Entomology and Molecular Biology, College of Life Sciences, Chongqing Normal University, Chongqing, PR China;* [2]*Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, UK;* [3]*Department of Entomology, The Natural History Museum, Cromwell Road, London, UK;* [4]*Department of Parasitology, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand;* [5]*Animal and Plant Sciences, The University of Sheffield, Western Bank, Sheffield, UK and* [6]*Faculty of Biological Sciences, The University of Leeds, Clarendon Way, Leeds, UK*

The effects of Pleistocene environmental fluctuations on the distribution and diversity of organisms in Southeast Asia are much less well known than in Europe and North America. In these regions, the combination of palaeoenvironmental reconstruction and inferences about population history from genetic data has been very powerful. In Southeast Asia, mosquitoes are good candidates for the genetic approach, with the added benefit that understanding the relative contributions of historical and current processes to population structure can inform management of vector species. Genetic variation among populations of *Anopheles minimus* was examined using 144 mtDNA COII sequences from 23 sites in China, Thailand and Vietnam. Haplotype diversity was high, with two distinct lineages that have a sequence divergence of over 2% and exhibit different geographical distributions. We compare alternative hypotheses concerning the origin of this pattern. The observed data deviate from the expectations based on a single-panmictic population with or without growth, or a stable but spatially structured population. However, they can be readily accommodated by a model of past fragmentation into eastern and western refugia, followed by growth and range expansion. This is consistent with the palaeoenvironmental reconstructions currently available for the region.
Heredity (2011) **106**, 241–252; doi:10.1038/hdy.2010.58; published online 2 June 2010

Introduction

The current distribution of genetic variation within a species is a product of its demographic history as well as the interaction between mutation, genetic drift and gene flow. It is notoriously difficult to distinguish historical effects from shorter-term consequences of dispersal patterns and limited population size but the distinction is important. Management of genetic diversity in endangered populations, prediction of the spread of insecticide resistance in agricultural pests or inferences about selection for local adaptation, for example, might be flawed if based on the assumption that current population structure is at an equilibrium between drift, mutation and dispersal. However, this assumption is often made. In these cases, historical effects may be considered a nuisance but in other contexts they are the main focus of interest. It is the imprint of historical changes on the distribution and abundance of organisms that is left in current population structure that underpins phylogeography (Avise, 2000). Often these spatial genetic patterns are on a much larger scale than the correlations generated by dispersal of individuals in stable populations.

In recent years, genetic signatures have been combined with palaeoenvironmental data to provide unprecedented insight into the post-glacial colonization of Europe and North America by a wide range of taxa (reviewed in Hewitt, 2004). We now have a clear picture of the locations of refugial areas and the variety of colonization routes followed. However, a natural tendency to see patterns in data, and to construct scenarios to explain those patterns, can easily lead to over-interpretation. This has led to calls for a more explicit statistical framework for phylogeographic analysis (Knowles and Maddison, 2002) and to controversy over the appropriate methodology. In particular, the use of the logical inference-based approach of nested clade phylogeographic analysis has been criticized for its high false-positive rate (Panchal and Beaumont, 2007). Nevertheless, it may have advantages over simulation-based model testing approaches that focus on a limited range of historical scenarios (Templeton, 2008, 2009).

Correspondence: *Professor RK Butlin, Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, South Yorkshire S10 2TN, UK.*
E-mail: r.k.butlin@sheffield.ac.uk
[7]*These authors contributed equally to this work.*

Whereas the impact of Pleistocene glaciations on North American and European biodiversity has been widely studied, much less attention has been given to other regions. Southeast Asia is rich in biodiversity and is experiencing rapid anthropogenic change as a result of economic development. There is, therefore, a need to understand the forces that have shaped the current distributions of species in the region, and the genetic variation within species. Palaeoenvironmental reconstructions for Southeast Asia do not have the firm foundations in numerous palynological data sets as they do in Europe and North America. However, the available palynological and fossil data suggest that during glacial periods much of the tropical forest of mainland Southeast Asia was replaced by pine and oak or savannah (Zheng and Lei, 1999; Penny, 2001; Hope *et al.*, 2004; White *et al.*, 2004). We suggest that mosquitoes can provide useful model systems for phylogeographic studies in Southeast Asia. A better taxonomic and ecological foundation exists than for many taxa in the region because mosquitoes are important as disease vectors. The vector status also means that there is a need to understand present-day distributions and population structure as an aid to control, for example in the management of insecticide resistance. There are many species, they are often abundant and they have a range of ecological requirements. Our previous study of *Anopheles dirus* and *Anopheles baimaii* (formerly known as *A. dirus* species A and species D, respectively) that overlap along the Thai-Myanmar border showed genetic signatures of population expansion (Walton *et al.*, 2000). Both species are heavily dependent on primary forest habitat (O'Loughlin *et al.*, 2008 and references therein). This, and the timing and geographic complexity of their population expansions, led to the suggestion that these expansions were in response to the post-glacial recovery and spread of tropical forests (O'Loughlin *et al.*, 2008).

Mosquitoes of the *Anopheles minimus* species complex (hereafter minimus complex) are widespread in hilly areas throughout the Oriental Region, extending northward to about 32° 30′ N in China, westward to Uttar Pradesh in India, southward through Thailand and into peninsular Malaysia and eastward to Taiwan and the Ryukyu archipelago of Japan (Chen *et al.*, 2002, 2006; Manguin *et al.*, 2008). Species of the complex are regarded as important vectors of human malaria throughout their distribution (Green *et al.*, 1990; Harrison *et al.*, 1990; Chen *et al.*, 2002; Garros *et al.*, 2006) showed that *A. minimus* consisted of two species in western Thailand, based on a lack of heterozygotes at two enzyme loci in sympatry, and informally designated them as species A and C. Sharpe *et al.* (1999) confirmed the presence of two species in western Thailand and established the use of the third domain (D3) of 28S ribosomal DNA as a standard for molecular identification. These species have since been formally named as *A. minimus* and *Anopheles harrisoni*, respectively (Harbach *et al.*, 2006, 2007), and this is the nomenclature used here. *A. minimus* is the predominant species of the minimus complex in Thailand and China (Green *et al.*, 1990; Chen *et al.*, 2002), and has also been recorded in Laos, Cambodia and Vietnam (Manguin *et al.*, 2008 and references therein). Species of the minimus complex are typically found in forested foothills where they use shaded stream margins as their larval habitat (Overgaard

et al., 2002). Although not found exclusively in forests, the abundance of species is generally positively associated with the extent of forest cover (Overgaard *et al.*, 2003; Rongnoparut *et al.*, 2005; Vanwambeke *et al.*, 2007), so they are expected to have experienced demographic change in response to forest contraction and fragmentation associated with Pleistocene climate change.

Currently, chemical measures remain central to the control of malarial parasites and their vectors. However, parasite and vector resistance to anti-malarial drugs and insecticides, respectively, is rapidly spreading in southern China and Southeast Asia (Lu, 1999). Population structure affects the potential for dispersal of insecticide resistance, and the potential of individuals to spread drug-resistant strains of malarial parasites. Therefore, an understanding of population structure is highly significant for malaria control (Collins *et al.*, 2000).

Because of its rapid evolution and maternal, non-recombining mode of inheritance, mitochondrial DNA (mtDNA) has been widely used to investigate population differentiation and evolutionary history of populations and subspecies (Avise, 2000). On the other hand, the mitochondrial genome is inherited as a single unit, independent of nuclear loci and with an effective population size one quarter that of autosomal loci. Its behaviour may, therefore, be atypical and an incomplete reflection of either population structure or history (reviewed in Ballard and Whitlock, 2004). This reservation must be kept in mind when making inferences based only on mtDNA data.

In this paper, we discuss sequence variation in the mtDNA COII gene within and among samples of *A. minimus* from southern China, Thailand and Vietnam. We consider the probability of obtaining the observed pattern of sequence variation under a range of increasingly complex historical demographic scenarios or current population structures and discuss the limits to interpretation of the data in the context of our current lack of knowledge about Southeast Asian environmental change. There is evidence for past fragmentation of this species' range followed by range expansion, with isolation by distance among current populations. Although this is broadly consistent with coarse-grained palaeoenvironmental reconstructions, many more data are needed for discriminating tests of alternative historical scenarios.

Materials and methods

Specimen collection

Mosquitoes were widely collected throughout southern China in 2000–2001, Thailand in 1996, 1998 and 2001, and northern Vietnam in 1998–1999. Most localities visited were areas with relatively high incidence of malaria and/or places where *A. minimus* s.l. had been collected previously by other researchers. Adults resting near or landing on humans and tethered bovines were captured with aspirators between 2000 and 2300 hours. Mosquitoes were killed by exposing them to ethyl acetate vapour, placed individually in Beem capsules (Agar Scientific Ltd, Stansted, Essex, UK) and stored in plastic bags with silica gel. Larvae and pupae were collected in streams and irrigation ditches at some sites and reared to adults. The keys of Harrison (1980) were used to identify

adults and larvae. These identifications were used to target specimens for further analysis, including molecular identification to differentiate cryptic species within the minimus complex.

Molecular identification and DNA sequencing
Mosquitoes were homogenized in buffer (0.01 M Tris, pH 7.8, 0.005 M EDTA, 0.5% SDS) and digested with $50\,\mu l\,ml^{-1}$ proteinase K at $37\,^{\circ}C$ overnight (Sambrook *et al*., 1989). DNA was then extracted with equal volumes of phenol/chloroform/isoamyl alcohol (25:24:1) twice and chloroform/isoamyl alcohol (24:1) once, followed by ethanol precipitation with 0.3 M sodium acetate and 2.0 volumes of 100% ethanol on ice for 60 min. After centrifugation at $13\,000\,g$ for 30 min and discarding the supernatant, the pellet was washed with 70% ethanol, dried and re-suspended in $20\,\mu l$ TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) before storage at $4\,^{\circ}C$. A negative control was included with every set of extractions.

The D3 region of the 28S rDNA was amplified for molecular identification, using primers D3a (5'-GACCCGTCTTGAAACACGGA-3', forward) and D3b (5'-TCGGAAGGAACCAGCTACTA-3', reverse), and the COII gene was amplified for population genetic study, using primers LEU (5'-TCTAATATGGCAGATTAGTGCA-3', forward) and LYS (5'-ACTTGCTTTCAGTCATCTAATG-3', reverse). As we found two deeply branched clades for the COII sequences (see below), which may represent cryptic species, we also sequenced 12 individuals for the ITS2 region using primers ITS2A (5'-TGTGAACTGCAGGACACAT-3', forward) and ITS2B (5'-TATGCTTAAATTCAGGGGGT-3', reverse) from Sharpe (1997). Amplifications were performed in $50\,\mu l$ volumes overlaid with two drops of mineral oil on a HYBAID OmniGene cycler (Thermo Hybaid, Ashford, UK). Each $50\,\mu l$ PCR reaction included 1/100 DNA of a whole mosquito, $5\,\mu l$ $10\times$ ReddyMix buffer (ABgene, Epsom, UK), $200\,\mu M$ dNTPs, 2 mM $MgCl_2$, each primer at $0.24\,\mu M$ and 1.3 units of Thermoprime Plus DNA Polymerase (ABgene). Reactions started with denaturation at $95\,^{\circ}C$ for 5 min, followed by 35 cycles, each cycle consisting of denaturation for 40 s at $95\,^{\circ}C$, annealing for 40 s at $55\,^{\circ}C$ and extension for 1 min at $72\,^{\circ}C$, with a final extension at $72\,^{\circ}C$ for 6 min. PCR products were subjected to electrophoresis through ethidium bromide-stained 1% agarose gels in $1\times$ TBE and visualized under UV light to check for successful amplification.

Molecular identification used the single-strand conformation polymorphism method of Sharpe *et al*. (1999) based on the D3 region. The COII gene of those individuals that were identified as *A. minimus* was sequenced, using the PCR primers, for the population genetics study. The D3 region of a smaller set of these individuals was sequenced to confirm the single-strand conformation polymorphism identifications. PCR products were cleaned using a spin column (Promega Wizard PCR Preps, Madison, WI, USA) and sequenced in both directions in an ABI 377 automated sequencer (PE Applied Biosystems, Warrington, UK). Sequences were edited manually and aligned using ClustalX (Thompson *et al*., 1997).

Analysis of mtDNA variation
The COII sequences were translated into amino acids using the *Drosophila* mtDNA genetic code in TRANSEQ (European Bioinformatics Institute: http://www.ebi.ac.uk/). Substitution models were fitted with TREE-PUZZLE (ver. 5.0; http://www.tree-puzzle.de). The parameters θ_π, based on the average pairwise number of differences between sequences, and θ_s, based on the number of segregating nucleotide sites per sequence, were used to estimate DNA polymorphism in populations or groups of populations in ARLEQUIN ver. 2.000 (Schneider *et al*., 2000). Population groups (Figure 1) were defined based on geographic proximity and regional topography. ARLEQUIN was also used to determine population genetic structure using analysis of molecular variance. Analysis of molecular variance estimates the proportion of total genetic variation attributable to different hierarchical levels based on the geographical distribution of haplotypes and takes into account the number of molecular differences through squared distances between haplotypes. The degree of inbreeding was estimated in terms of Φ-statistics in a manner analogous to F-statistics. As ARLEQUIN does not accommodate the HKY evolutionary model suggested by TREE-PUZZLE, we used the less complex K-80 model for these comparisons. The significance of the Φ-statistic was tested by permutation of the haplotypes among populations and groups. Genetic isolation by geographic surface distance was assessed by Mantel test for collection localities with sample sizes of five or higher.

The frequency distribution of the numbers of segregating sites in all possible pairwise comparisons, known as a mismatch distribution, was calculated and compared with the expectation under the stepwise expansion model of population demography in ARLEQUIN (Schneider *et al*., 2000).

'Tests of neutrality' have been designed to detect deviations from the pattern of polymorphism expected from neutral evolution in a random mating population of constant size. Tajima's test statistic D, which compares the two mutation parameter estimators, θ_π and θ_s, and Fu's F_s statistic were used, both based on the infinite-sites model without recombination, and thus appropriate for short mtDNA sequences. These statistics are sensitive to different aspects of the pattern of variation among haplotypes and can be used as summaries of the pattern, as well as tests of departure from an expected pattern of a stable population. All of these analyses were carried out in ARLEQUIN.

Genealogical relationships among haplotypes were constructed by statistical parsimony and Nested Clade Analysis (Templeton, 1998) was conducted using the automated analysis and inference protocol of Panchal and Beaumont (2007). As *A. minimus* is expected to have had a continuous distribution before recent, extensive deforestation, we emulated this with a 2.5-degree grid of occupied but unsampled localities. Each location was considered to have an area of 150 km radius that overlapped with neighbouring locations.

The LAMARC suite of programs version 1.1 (Kuhner, 2006) was used to estimate effective population size and population growth rate or migration rate using a maximum-likelihood method. SIMCOAL (ver. 1.0) was used to generate 1000 simulated data sets under various demographic models for comparison with the observed data (Excoffier *et al*., 2000). The sum of squared deviations (SSDs) was used to compare the observed mismatch
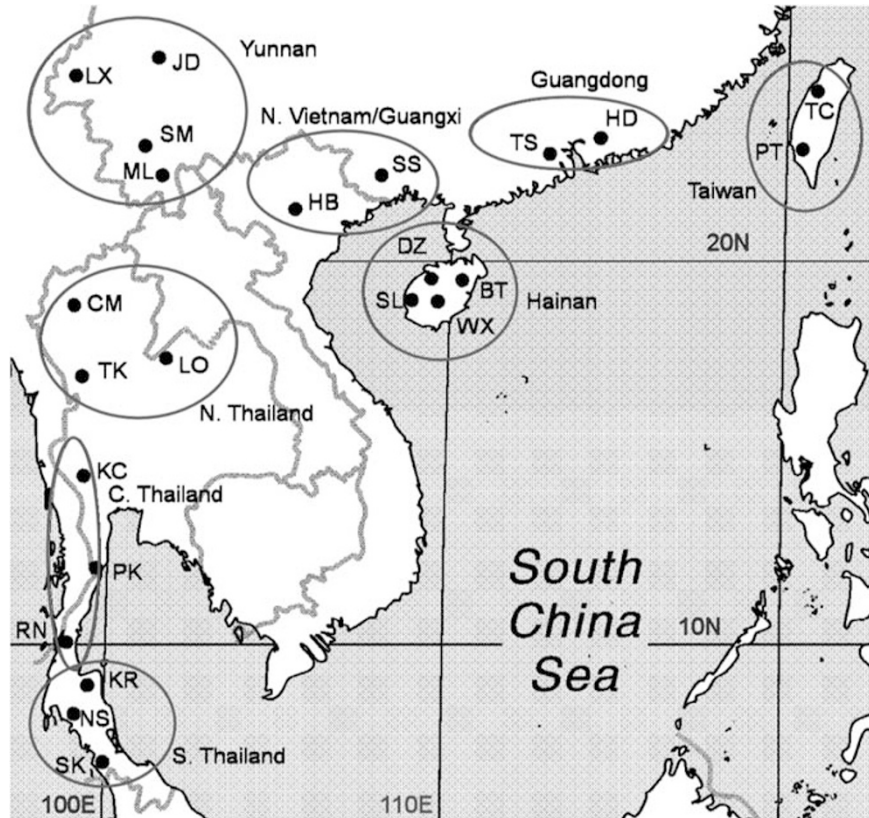
Figure 1 Collecting localities for populations of *Anopheles minimus* in China, Vietnam and Thailand. The 23 populations were divided into nine population groups (circled) based on their geographical distributions.

distribution to mismatch distributions from simulated data sets in a manner analogous to the test of the stepwise expansion model implemented in ARLEQUIN. In ARLEQUIN, observed and simulated distributions are compared with the theoretical expectation. Here, we compared the mean SSD between observed and simulated distributions with the range of mean SSDs between any one simulated distribution and the remaining 999.

Results

Approximately 1300 of 5000 *Anopheles* specimens collected in southern China, Thailand and northern Vietnam were found to be species of the Myzomyia Series of the subgenus *Cellia*, and ~600 specimens were morphologically identified to the miminus complex (Harrison, 1980). DNA was successfully extracted from these samples, and PCR products of the D3 region were run on single-strand conformation polymorphism gels (Sharpe *et al.*, 1999). As a result, some 250 samples were unambiguously identified as *A. minimus*, and sequencing in the D3 region of 45 randomly selected samples confirmed the single-strand conformation polymorphism identification. As there were many *A. minimus* specimens from some sites, we selected 144 individuals from 23 sites (Table 1; Figure 1) for sequencing of the COII gene for the phylogeographic study.

COII sequence polymorphism
Sequences were trimmed to a consistent length of 685 bp and their alignment was unambiguous. The sequences were very AT rich (75.94% A and T bases). Nucleotide substitutions were identified at 54 of the 685 sites, of which 43 were transitions (79.6%), eight were transversions (14.8%) and three (at positions 303, 321 and 369) had three segregating bases (Supplementary Information Table S1). Fourteen substitutions at six sites resulted in amino-acid changes (directionality based on comparison with other members of the minimus complex): at site 34 (S→G) for haplotypes 7, 8, 35 and 53; at site 84 (L→F) for haplotype 35; at site 368 (I→M) for haplotype 8 and (I→T) for haplotype 28; at site 389 (T→M) for haplotypes 8 and 9; at site 493 (V→I) for haplotypes 10, 11 and 36; and at site 550 (F→V) for haplotypes 6 and 65. Sequence differences among these haplotypes are shown in Supplementary Table S1, and haplotypes have been deposited in EMBL with accession numbers FN433526–FN433595.

The distribution of haplotypes among the 23 populations is shown in Table 1. Haplotype diversity was very high, with 70 haplotypes found among 144 individuals. This is in accord with the generally high levels of nucleotide diversity estimated by θ_π and θ_s.

Genealogical relationships among haplotypes
The haplotype network generated by TCS shows two widely separated networks (hereafter Lineage A and Lineage B, Figure 2). No connection was found between these two lineages with the 95% probability of statistical parsimony criterion. A neighbour-joining tree for the 70 haplotypes rooted with sequences from *A. harrisoni*

Table 1 Data summary for populations, haplotypes and nucleotide diversity of *An. minimus*

Population group and collection sites	Code	N	Haplotypes[a]	Lineages A+B				Lineage A				Lineage B			
				h (s.d.)	S	θ_π	θ_s	h (s.d.)	S	θ_π	θ_s	h (s.d.)	S	θ_π	θ_s
China: Guangdong															
Huidong: Daling	HD	7	**1**(3), 33, 36, 37(2)					0.857 (0.108)	9	0.004	0.005				
Taishan: Nongwen	TS	1	**49**												
China: Hainan															
Baoting: Nanling	BT	1	**37**					0.804 (0.091)	10	0.003	0.004				
Changjiang: Shilu	SL	8	**1**(4), **10**(2), **43**, 48												
Changjiang: Wangxia	WX	8	**1**(3), **17**, 21, **37**, 42, 44												
Danzhou: Bayi	DZ	1	**1**												
China: Taiwan															
Taichung	TC	7	**1**(2), **37**(2), **47**(2), 51					0.714 (0.116)	11	0.006	0.005	—	—	—	—
Pingtung	PT	8	40, 41, **47**(6)												
Vietnam (Hoa Binh) and China (Guangxi)															
Tan Lac: Phu Cuong	HB	10	13, **20**(4), **38**(3), **39**(2)					0.778 (0.091)	6	0.004	0.003	—	—	—	—
Shangsi: Zhaoan	SS	1	**31**												
China: Yunnan															
Jingdong: Chengjiao	JD	13	7, 8, **9**, 14, **16**, 24, **31**, 32, [54, 55, **58**, 65, 70]	0.987 (0.011)	39	0.017	0.014	0.971 (0.032)	19	0.006	0.008	0.975 (0.029)	17	0.005	0.007
Luxi: Wanding	LX	2	**2**, [62]												
Mengla: Menglun	ML	15	**2**, **16**, 29, 35, **50**, [**52**(2), 53, 56(2), 63, **64**(2), 67, 68]												
Simao: Cuiyun	SM	2	6, 26												
Northern Thailand															
Chiang Mai	CM	3	30, **38**, **49**	0.974 (0.029)	27	0.016	0.011	0.985 (0.040)	15	0.007	0.007	0.800 (0.172)	5	0.003	0.003
Loei	LO	12	**1**(2), 3, 15, 23, **25**, 34, **45**, [**52**(2), 58, 66]												
Tak	TK	5	**2**(3), [**52**, 59]												
Central Thailand															
Kanchanaburi	KC	8	**1**, **2**, 4, 19, 22(2), 46, [**64**]	0.956 (0.045)	28	0.010	0.013	0.949 (0.051)	17	0.007	0.008	—	—	—	—
Prachuap Khiri Khan	PK	3	**25**, 27, 28												
Ranong	RN	1	**50**												
South Thailand															
Krabi	KR	12	**1**(2), **2**, 12, **17**(2), **37**, **43**(2), **45**(3)	0.943 (0.023)	25	0.014	0.009	0.918 (0.036)	11	0.004	0.005	0.786 (0.151)	4	0.002	0.002
Nakhon Si Thammarat	NS	13	**2**, 5, 11, **17**, 18, [**52**(4), 57, 60, 61, 69]												
Songkhla	SK	3	**2**(2), **17**												

h is haplotype diversity, *S* is the number of segregating sites and θ_π and θ_s are the estimates of nucleotide diversity.
[a]The numbers in parentheses indicate the frequency of the haplotype, and the haplotypes in square brackets belong to Lineage B. Haplotypes in bold occur more than once.

resulted in a topology with two deeply separated branches corresponding to Lineages A and B. The average pairwise sequence difference between the two lineages was 2.2% and a minimum of 11 substitutions was needed to connect them. The mismatch distribution was strongly bimodal, with each mode smooth (Figure 3). The haplotypes in Lineage A were represented in all populations investigated but those in Lineage B were restricted to the western part of the sampled area (roughly west of longitude 101° E; Table 1; Figure 1).

It is possible that the two mitochondrial lineages represent cryptic species. Therefore, we sequenced 512bp of ITS2 from eight individuals belonging to Lineage A (from sites TC, HD, WX, SS, NS, KC, HB (see Table 1) and Linxiang, Yunnan) and four individuals from Lineage B (from sites ML, JD, LO and NS). All sequences were identical, regardless of lineage. A representative sequence has been submitted to EMBL with accession number FN646403.

Comparisons among demographic models
The occurrence of two distinct groups of haplotypes might be taken to suggest past population fragmentation,
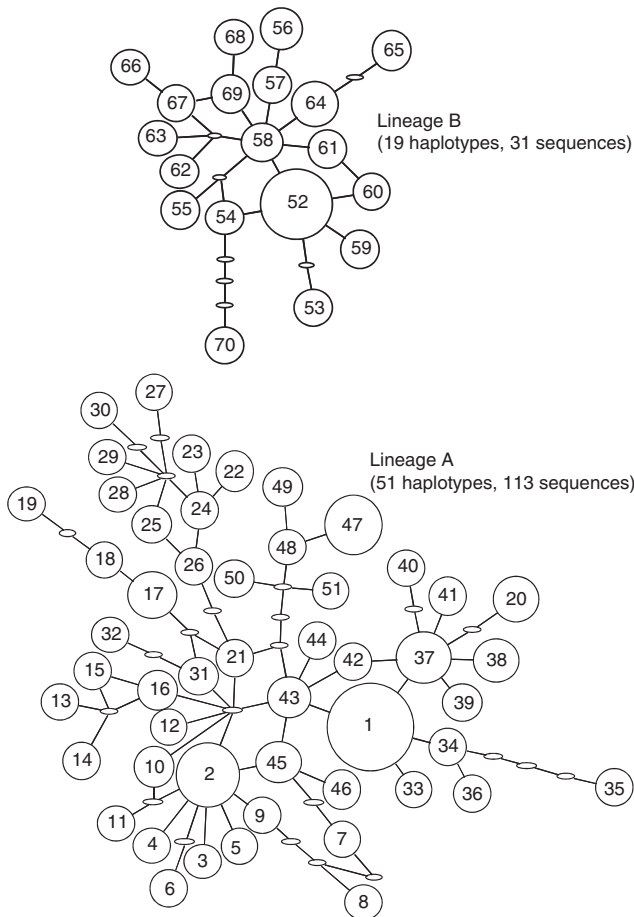
Figure 2 Two separate haplotype networks for *Anopheles minimus*. The size of a circle corresponds to the haplotype frequency, and a unit branch represents one mutation. Small ovals indicate unobserved, inferred haplotypes.
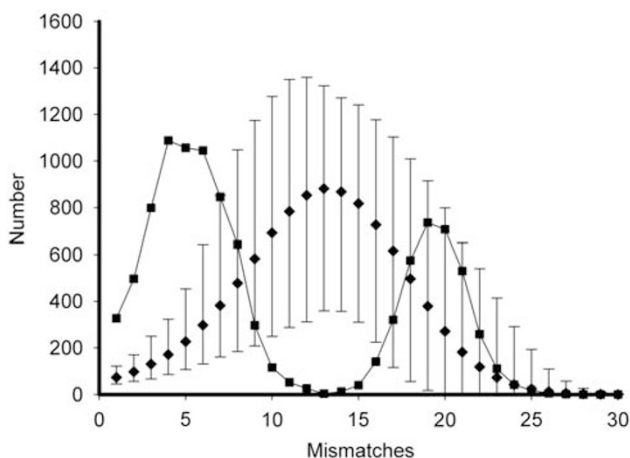

Figure 3 Observed mismatch distributions among haplotypes in Lineages A + B (squares) with the results of SIMCOAL runs using the parameters inferred by FLUCTUATE for a single, exponentially expanding population (means, diamonds and 95% confidence intervals from 1000 simulations).

perhaps during the last glaciation, followed by range expansion and overlap. A nested clade phylogeographic analysis, based on one resolution of the homoplasious

connections in the haplotype network shown in Figure 2, using the criteria of Pfenninger and Posada (2002), revealed significant signals of past fragmentation at the highest nesting level, contiguous range expansion at the next and some lower levels, and isolation-by-distance in one zero-step and one one-step clade (results available on request). Some long-distance dispersal events were inferred across gaps in the distribution that are currently sea (between the mainland and the islands of Taiwan and Hainan). However, during glacial periods the Sunda shelf, an extension of the continental shelf of Southeast Asia that is currently under sea, was exposed (Voris, 2000). Overall, these inferences are consistent with expectations based on Pleistocene climatic and environmental perturbation. However, simpler population histories might also generate such a pattern and nested clade phylogeographic analysis has been criticized for failing to make appropriate statistical comparisons among competing historical demographic scenarios and for ignoring the uncertainty in the genealogy of haplotypes (Knowles and Maddison, 2002), requiring the resolution of loops in the haplotype network without strong justification and for generating high false-positive rates (Panchal and Beaumont, 2007) (but see Templeton, 2008, 2009) for counter arguments). Genealogical uncertainty is particularly problematic here because the strong AT bias and high variance in substitution rate lead to high levels of homoplasy. The significant demographic signals in this case were not stable when alternative resolutions of the homoplasious connections in the network were investigated.

In testing the alternative phylogeographic hypotheses, it is also important to take into account the effects of coalescent stochasticity (Carstens *et al.*, 2005). In particular, the time to the most recent common ancestor of a sample of lineages is highly variable even in a single population of constant size (Rosenberg and Nordborg, 2002). The time from the penultimate to the last coalescent event is particularly variable and so genealogies with deep central branches are common. They may be more common relative to the lengths of the terminal branches under some forms of population expansion. Population structure also tends to generate deep central branches combined with partial or even complete geographic separation (Irwin, 2002). Therefore, we have compared the plausibility of different demographic hypotheses, given our observed sample of 144 haplotypes.

First, we used the program FLUCTUATE in the LAMARC suite to test for population growth. FLUCTUATE makes simultaneous, coalescent estimates of the present-day value of θ, θ_0 and the population growth rate, g, assuming an exponential model of growth and using a maximum-likelihood approach (Kuhner, 2006). The algorithm in FLUCTUATE is sensitive to the initial values of θ and g, and the initial genealogical tree of sequences. Therefore, we first ran 10 short chains of 1000 steps and two long chains of 20 000 steps with Watterson's estimate of θ, a g-value close to zero and UPGMA tree of sequences as initials. We also compared the output with runs using random starting trees. We used the substitution model parameters determined in TREE-PUZZLE. We then ran 20 short chains of 1000 steps and five long chains of 20 000 steps using the estimates of θ and g and the best tree (a good approximation of the

Table 2 Estimates of effective population size and exponential growth rate from FLUCTUATE and parameters of the stepwise expansion model from ARLEQUIN

Exponential growth	$\theta \pm s.d.$	N_0	$g \pm s.d.$
Lineages A+B	0.2644 ± 0.0156 (0.261)	13.2×10^6	305.4 ± 14.3 (307.0)
Lineage A	0.1584 ± 0.0154 (0.291)	9.92×10^6	808.2 ± 68.3 (474.7)
Lineage B	0.1289 ± 0.0224 (0.224)	6.45×10^6	1942 ± 103 (1863.9)
Stepwise growth	τ (95% CI)	θ_0 (95% CI)	θ_1 (95% CI)
Lineages A+B	0.00250 (0.00031, 0.0215)	0.00987 (0, 0.0254)	0.0775 (0.0204, 6.676)
Lineage A	0.00714 (0.00387, 0.0102)	3×10^{-6} (0, 0.00293)	0.0361 (0.0166, 9.708)
Lineage B	0.00415 (0.00200, 0.00549)	0 (0, 0.00222)	0.332 (0.0149, 11.93)

Abbreviation: CI, confidence interval.
For the exponential model, point estimates of θ and g are shown with approximate s.d. using the UPGMA starting tree (mean from five random starting trees in parenthesis). The exponential growth rate, g, is in units of μ^{-1} per generation. N_0, the current effective population size of females, is estimated from θ ($\theta = 2N\mu$, where N is the effective number of females) assuming μ of 1×10^{-8} per year (10^{-9} per generation, 10 generations per year) (Powell *et al.*, 1986). For the stepwise model, τ is the time since the increase in population size, scaled by the mutation rate ($\tau = 2\mu t$ generations); θ_0 and θ_1 are proportional to the population size before and after expansion, respectively.

maximum-likelihood tree) produced in the first computation as initials to make the final estimates of θ and g. Significant population growth was detected ($\Delta \ln L = 21.45$), but the estimated rate of growth was low (Table 2).

Although FLUCTUATE indicates that the data fit the exponential growth model better than a stable population-size model, it does not necessarily follow that they are a good fit to the data. We assessed the extent to which the demographic parameters estimated by FLUCTUATE could explain the features of the observed data by generating 1000 simulated data sets using SIMCOAL. For each simulated data set, we used ARLEQUIN to generate a mismatch distribution, Tajima's D and Fu's F_s. We then compared the observed mismatch distribution and neutrality statistics with the range of simulated values.

Simulations of a single exponentially growing population, using the FLUCTUATE maximum-likelihood estimate of growth and present population size, never produce mismatch distributions with the strongly bimodal pattern seen in our data (Figure 3). The mean SSD between observed and simulated data fell outside the range of comparisons among simulated distributions. Although both D and F_s fall within the range of observed values, simulations never produced the combination of strongly negative F_s and near zero D observed in our data (Figure 4a). F_s is more sensitive to population growth than is D (Fu, 1997) and, therefore, this result might indicate either a rapid or a recent expansion or an effect of population structure. Alternatively, departures from the infinite sites model might influence the two statistics differently, and the substitution model used in
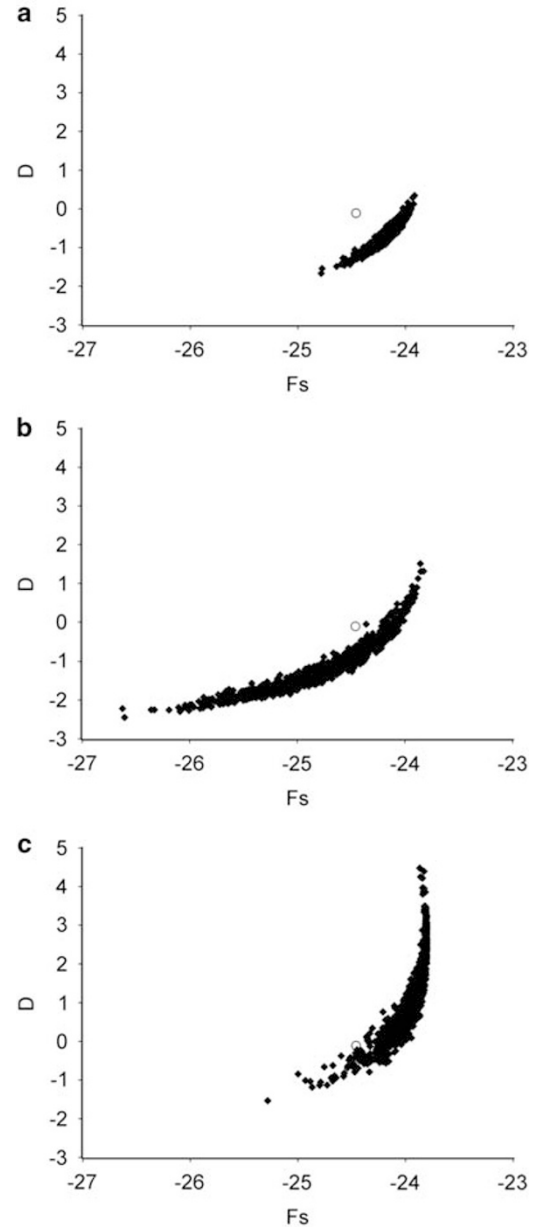

Figure 4 Comparisons between the observed D and F_s estimates (open circle) with those obtained from 1000 SIMCOAL simulated data sets (filled diamonds) under each of three scenarios: an exponentially expanding population (**a**), a population that has experienced a stepwise expansion (**b**) and a structured population of constant size (**c**). See text for further explanation.

the simulations might not fully account for the pattern in the data.

We also tested the stepwise expansion model for a single population, a scenario approximating logistic growth with instantaneous increase of population size. ARLEQUIN provides estimates of the parameters of this model and tests the fit to the model by parametric bootstrap (Schneider *et al.*, 2000). The estimated values (95% confidence intervals) were: $\tau = 0.0025$ (0.00031, 0.0215), $\theta_0 = 0.01$ (1.5×10^{-6}, 0.027) and $\theta_1 = 0.24$ (0.026, 8.07). With a mutation rate of 10^{-9} per base per generation, this corresponds to an expansion 125 000 years ago (assuming 10 generations per year) from a

Table 3 Hierarchical analysis of molecular variance for Lineages A and B

Lineage	Variance components	d. f.[a]	Sum of squares	Variance component	Φ-statistics	P
Lineage A	Among groups	7	82.42	0.433	$\Phi_{CT} = 0.11$	<0.01
	Among populations/groups	15	66.56	0.398	$\Phi_{SC} = 0.12$	<0.001
	Within populations	90	263.63	2.929		
Lineage B	Among groups	2	3.03	−0.295	$\Phi_{CT} = -$	
	Among populations/groups	4	12.29	0.434	$\Phi_{SC} = 0.22$	NS
	Within populations	24	38.75	1.615		
Lineage B, no groups	Among populations	6	10.68	0.132	$\Phi_{ST} = 0.10$	NS
	Within populations	24	29.92	1.247		

[a]Degrees of freedom.

population size of ∼5 to ∼40 million females. This model was not rejected on the basis of the SSD of the mismatch distribution from the expectation. When we simulated sudden expansion using these parameters in SIMCOAL, strongly bimodal mismatch distributions similar to the observed distribution were sometimes produced because divergent lineages can survive from the pre-expansion population and leave multiple descendent haplotypes. The mean SSD for the observed distribution was exceeded by 118/1000 simulations. The observed D and F_s values fall just outside the range of simulated combinations (Figure 4b) but, of course, this range would be larger if uncertainty in the expansion parameters was taken into account.

Analysis of molecular variance in Arlequin indicates substantial population structure at the 'population group' and 'among populations within group' levels in Lineage A (Table 3; groups as indicated in Figure 1). Although not significant, it seems likely that a similar level of genetic structure is present in Lineage B but cannot be detected here due to smaller sample sizes and restricted sampling area. The Migrate package, with the N island model of migration, estimated island population size at 24 000 (19 000, 34 100) and the number of migrants at 4.5 (3.6, 6.8). There was a significant correlation between genetic distance (corrected mean pairwise difference) and geographic distance (Mantel test: $r = 0.22$, $P = 0.0027$). This pattern is partly due to the different distributions of the A and B lineages, with Lineage B restricted to the western part of the study area (Table 1). Geographic structure is not expected under the preceding demographic models, which assume panmixia. As population structure can increase the time to the most recent common ancestor and generate deeply branched lineages with restricted ranges (Irwin, 2002), this structure might account for the observed sets of divergent haplotypes.

SIMCOAL was used to simulate samples from a structured population using the migration and population size parameters estimated by Migrate. The observed D and F_s values fall just within the range of simulated values (Figure 4c). The mismatch distributions were frequently bimodal but rarely had as high a peak as the observed distribution (Figure 3) at low numbers of mismatches (3–6, exceeded ⩽12 times in 1000 simulations). The mean SSD for the observed data was exceeded only 13/1000 times by the simulated data sets.

Finally, a model of fragmentation during the last glaciation can be considered, as suggested by our original nested clade analysis. As the last glaciation began ∼100 000 years (10^6 mosquito generations) ago, and is expected to have resulted in periods of reduced population size, it is likely to have generated two or more independent mitochondrial clades reflecting distinct refugia. This suggests that Lineages A and B can be considered separately, representing the admixture of populations that had expanded from eastern and western refugia. FLUCTUATE detected significant population growth in each case (Table 2), and the inferred rate of growth was greater than for the combined lineages. The stepwise expansion model was not rejected either and, similarly, it suggests large expansions in both lineages (Table 2). The estimates for all parameters are similar for the two lineages. This implies that Lineage B has a wider distribution than the area sampled here.

SIMCOAL, with parameters from the exponential expansion model, generates unimodal distributions encompassing the observed values. For Lineage B, the observed combination of D and F_s falls within the range of simulated values but this is not true for Lineage A. This may be because there is significant population structure within Lineage A at both group and population levels (Table 3), but not within Lineage B. Isolation by distance also remains significant within Lineage A ($r = 0.26$, $P < 0.01$).

Patterns of genetic differentiation within Lineage A

Although there is no spatial structure among the small number of localities sampled within Lineage B, there are patterns of spatial variation within Lineage A. Samples from Yunnan, northern, central and southern Thailand are all similar (Table 4). These are the western groups where Lineage B is also present. The Vietnam, Taiwan and Hainan + Guangdong groups are all distinct from one another and from the western region groups. The most diverse samples were from the western region (ML, LO, KC and NS; Table 1) and at the group level central Thailand, Yunnan and northern Thailand, respectively, had the highest diversity. These patterns suggest a history of colonization from west to east, but the absence of samples from the potentially critical areas of Cambodia and southern Vietnam makes any firm inference premature.

Discussion

Demographic history

When the two lineages were considered jointly as a single genealogy, the FLUCTUATE analysis indicated

Table 4 Genetic distances between population groups of Lineage A

Group	Yunnan	Hainan	Guangdong	Taiwan	Vietnam	N Thailand	C Thailand
Yunnan							
Hainan	0.172*						
Guangdong	0.208*	0.007					
Taiwan	0.325*	0.290*	0.238*				
Vietnam	0.275*	0.256*	0.233*	0.394*			
N Thailand	−0.014	0.090*	0.113*	0.283*	0.231*		
C Thailand	0.008	0.122*	0.121*	0.219*	0.204*	−0.021	
S Thailand	0.056	0.134*	0.217*	0.352*	0.312*	0.027	0.054

The genetic distances between are presented as pairwise F_{ST} values. *$P < 0.05$.

that the data were a significantly better fit to a demographic model of expansion rather than one of stability, indicating recent population growth in this species. Although bottlenecking is expected to increase the possibility of divergent lineages being generated (due to the loss of intermediate haplotypes by genetic drift), simulations of expanding populations rarely produced the deeply divergent lineages seen in our data. Although other simulations have shown that it is possible for deeply divided genealogies to evolve even when there are no barriers to gene flow (Irwin, 2002), this pattern was produced primarily for organisms with very low population densities, which is not a characteristic of *A. minimus*. Although the deeply bifurcating genealogy we sampled could have occurred by chance in a single stable or expanding population, our analyses show that it is more likely that other factors are responsible.

One such factor is population structure. Simulations have shown that range expansions associated with low *Nm* values (where *N* is the effective population size and *m* is the migration rate) can have bimodal mismatch distributions (Ray *et al.*, 2003). However, in this case, the lower mode is due to a large number of very low (often zero) pairwise differences between haplotypes due to coalescence within the same deme, whereas in our data, the mismatch peak with the lowest mean number of pairwise differences involves comparisons between haplotypes of the same lineage, largely from different populations. Further model testing indicated that population structure was unlikely to explain the deeply branched genealogy, and that this pattern was best explained by allopatric fragmentation. We therefore conclude that our data support the historical fragmentation scenario, even in the absence of an *a priori* expectation for such a history. Our systematic comparison of a set of increasingly complex models, and comparisons of the data to simulations using multiple measures to assess the fit, make this conclusion relatively secure. Nevertheless, it must always be the case that other scenarios exist and some are capable of generating the observed patterns cf. (Templeton, 2008). It also remains true that our analyses are based on a single marker, mtDNA, which may not be representative of the genome as a whole and the analyses ignore the possible impact of selection. Although most of the polymorphism in mtDNA sequences is likely to be neutral, positive or negative selection on any site within the mtDNA genome will influence the pattern of variation across the whole molecule (Ballard and Whitlock, 2004). Therefore, the historical scenario we propose needs to be tested with other markers and/or other species.

Table 5 Diversity statistics and tests of neutrality

θ Values and tests	Lineages A+B	Lineage A	Lineage B
θ_s	9.741	7.358	5.006
θ_π	9.045	4.324	2.692
Tajima's D	−0.220	−1.269*	−1.600*
Fu's F_s	−24.450**	−25.704**	−14.235**

*$P < 0.1$.
**$P << 0.01$.

Lineage distribution

The distribution of the minimus complex is thought to extend north-westward to Uttar Pradesh in India and southward into peninsular Malaysia (Harrison *et al.*, 1990; Chen *et al.*, 2002, 2006; Harbach *et al.*, 2007), regions that we were not able to sample in this study. Although the majority of samples here belonged to Lineage A, Lineage B had only slightly lower genetic diversity implying similar population sizes, and suggesting that we have not sampled a significant part of its range. As Lineage B is found mixed with Lineage A in western Thailand and western China, it is likely that the range of Lineage B extends westwards through Myanmar and potentially into Bangladesh and India. Therefore, further sampling in western Southeast Asia and possibly also into peninsular Malaysia is needed to firmly establish the geographic distributions of these lineages.

Biogeography

When analysed separately, Lineages A and B each show evidence of recent population expansion from the fit to exponential and stepwise population growth models (Table 2), significantly negative neutrality tests (Table 5) and smooth unimodal mismatch distributions (data not shown). On the basis of the stepwise expansion model, expansion in Lineage A dates to ∼357 000 years ago and in Lineage B to ∼208 000 years ago. As the confidence intervals of these estimates overlap, the expansions could have occurred at the same time. Other forest insect species from Indochina show similarly dated genetic signals of expansion for mtDNA: *A. dirus* (O'Loughlin *et al.*, 2008), *Simulium tani* blackflies (Pramual *et al.*, 2005) and *Anopheles jeyporiensis* (Chen *et al.*, 2004). Stone oaks in mainland (but not insular) Southeast Asia also have a star-like chloroplast DNA genealogy characteristic of demographic expansion (Cannon and Manos, 2003). Signals of expansion from multiple forest species are most likely to be a response to a common factor. As we have suggested previously, this factor seems most likely

to be the spread of forests from refugial areas following the onset of an interglacial period (Chen *et al.*, 2004; Pramual *et al.*, 2005; O'Loughlin *et al.*, 2008). The divergence of Lineages A and B, which dates to ~1 mya using the conventional calibration rate for animal mtDNA of 2% divergence per million years (Powell *et al.*, 1986), has likely accumulated over multiple glacial cycles. Our findings therefore concur with larger-scale phylogeographic studies on tigers (Luo *et al.*, 2004), dholes (Iyengar *et al.*, 2005) and elephants (Fleischer *et al.*, 2001) that Pleistocene climatic change has played a major role in shaping the genetic diversity of the mainland Southeast Asian biota.

The different distributions of the two lineages suggest isolation in allopatric eastern and western forest refugia. This is broadly consistent with coarse-grained palaeo-environmental models of the LGM, which show eastern and western tropical forest may have survived along the Tenasserim Mountains on the Thai-Myanmar Border and the Annamite mountain range running from northern to southern Vietnam (Ray and Adams, 2001). During cooler, drier periods, forest refugia are most likely to survive in the mesic valleys of such mountainous regions (Brandon-Jones, 1996). The higher diversity of Lineage A in the west of its sampled distribution suggests foothills to the east of mountains in Yunnan and/or northern/central western Thailand as a possible refugial area. The western foothills of the same region could be the refugial area of Lineage B (if the peaks of the Tenasserim Mountains formed a barrier to gene flow during glacial periods) or this could be much further west. These propositions represent hypotheses that need to be tested using further sampling and additional genetic markers.

It is likely that *A. minimus*, along with other forest-associated mosquitoes, has been impacted by human activities, especially deforestation but also malaria-vector control strategies. Analysis of short mtDNA sequences such as ours is unlikely to have the resolution to detect signals of population decline from such impacts.

Are the lineages species?

The 2.2% level of pairwise divergence between the two *A. minimus* lineages is not much less than that between each lineage and the closely related species *A. harrisoni* (2.6 and 2.7% between *A. harrisoni* and Lineages A and B, respectively). The estimated divergence time of 1 million years could be sufficient for the lineages to have formed distinct species, based on estimates of 1–2 million years for allopatric speciation in *Drosophila* but probably less than this if there is a phase of sympatry or ecological adaptation involved (Coyne and Orr, 2004). Behavioural differences have also been noted between *A. minimus* populations in eastern Southeast Asia (Van Bortel *et al.*, 2004). Although this might be suggestive of the presence of more than one species within the taxon, the differences appear to be north-south, so are not obviously consistent with the east-west distribution of the lineages.

Ribosomal DNA is often used in *Anopheles* to distinguish isomorphic species as between-species differences tend to be rapidly fixed for this marker (Walton *et al.*, 1999). At the rDNA D3 region, *A. minimus* and its close relatives differ from each other by at most 5 bp (*A. minimus* and *A. harrisoni*) but by as little as 1 bp between *Anopheles fluviatilis* species T and U (Chen *et al.*, 2002).

The ITS2 region evolves more rapidly and often discriminates between very closely related *Anopheles* taxa (Beebe *et al.*, 1999; Walton *et al.*, 2007). Therefore, the lack of differentiation for the D3 and ITS2 regions argues against the presence of distinct species. Allozyme surveys in the past did not give any indication of cryptic species within *A. minimus* (then known as *A. minimus* A; Green *et al.*, 1990). However, we cannot exclude the possibility that the lineages represent two species that have not yet accumulated differences at D3 or ITS2.

Conclusion

Use of nuclear loci would clearly aid in the interpretation of the demographic history and current species status of the two *A. minimus* mtDNA lineages. However, understanding the biogeographic history of Southeast Asia, and specifically testing the hypothesis of an east-west suture zone, might be advanced more rapidly by making comparisons with the distribution of mtDNA genetic diversity in other co-distributed forest-associated species, including other mosquito taxa. Determining the patterns and processes of the diversification of forest fauna in mainland Southeast Asia is important for understanding the origins of tropical biodiversity and for identifying conservation priorities in this region (Moritz *et al.*, 2000).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

References

Avise JC (2000). *Phylogeography: the History and Formation of Species*. Harvard University Press: Cambridge, MA.

Ballard JWO, Whitlock MC (2004). The incomplete natural history of mitochondria. *Mol Ecol* **13**: 729–744.

Beebe NW, Ellis JT, Cooper RD, Saul A (1999). Nucleotide DNA sequence analysis of the ribosomal DNA ITS2 region for the *Anopheles punctulatus* group of mosquitoes. *Insect Mol Biol* **8**: 381–390.

Brandon-Jones D (1996). The Asian Colobinae (Mammalia: Cercopithecidae) as indicators of Quaternary climatic change. *Biol J Linn Soc* **59**: 327–350.

Cannon CH, Manos PS (2003). Phylogeography of the Southeast Asian stone oaks (Lithocarpus). *J Biogeog* **30**: 211–226.

Carstens BC, Degenhardt JD, Stevenson AL, Sullivan J (2005). Accounting for coalescent stochasticity in testing phylogeographical hypotheses: modelling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. *Mol Ecol* **14**: 255–265.

Chen B, Harbach RE, Butlin RK (2002). Molecular and morphological studies on the *Anopheles minimus* group of mosquitoes in southern China: taxonomic review,

distribution and malaria vector status. *Med Vet Entomol* **16**: 253–265.

Chen B, Harbach RE, Butlin RK (2004). Genetic variation and population structure of the mosquito *Anopheles jeyporiensis* in southern China. *Mol Ecol* **13**: 3051–3056.

Chen B, Butlin RK, Pedro PM, Wang XZ, Harbach RE (2006). Molecular variation, systematics and distribution of the *Anopheles fluviatilis* conplex in southern Asia. *Med Vet Entomol* **20**: 33–43.

Collins FH, Kamau L, Ranson HA, Vulule JM (2000). Molecular entomology and prospects for malaria control. *B World Health Organ* **78**: 1412–1423.

Coyne JA, Orr HA (2004) Speciation, Chapter 12, Speciation and Macroevolution. Sinauer Associates Inc.: Sunderland, MS.

Excoffier L, Novembre J, Schneider S (2000). SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered* **9**: 506–509.

Fleischer RC, Perry EA, Muralidharan K, Stevens EE, Wemmer CM (2001). Phylogeography of the Asian elephant (*Elephas maximus*) based on mitochondrial DNA. *Evolution* **55**: 1882–1892.

Fu YX (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

Garros C, Van Bortel W, Trung HD, Coosemans M, Manguin S (2006). Review of the Minimus Complex of *Anopheles*, main malaria vector in Southeast Asia: from taxonomic issues to vector control strategies. *Trop Med Int Health* **11**: 102–114.

Green CA, Gass RF, Munstermann LE, Baimai V (1990). Population genetic evidence for two species in *Anopheles minimus* in Thailand. *Med Vet Entomol* **4**: 25–34.

Harbach RE, Garros C, Manh ND, Manguin S (2007). Formal taxonomy of species C of the *Anopheles minimus* sibling species complex (Diptera: Culicidae). *Zootaxa* **1654**: 41–54.

Harbach RE, Parkin E, Chen B, Butlin RK (2006). *Anopheles* (*Cellia*) *minimus* Theobald (Diptera: Culicidae): neotype designation, characterization, and systematics. *P Entomol Soc Wash* **108**: 198–209.

Harrison BA (1980). Medical entomology studies XIII The Myzomyia series of Anopheles (Cellia) in Thailand, with emphasis on intra-interspecific variations (Diptera: Culicidae). *Contrib Am Entomol Inst* **17**: 1–195.

Harrison BA, Rattanarithikul R, Peyton EL, Mongkolpanya K (1990). Taxonomic changes, revised occurrence records and notes on the Culicidae of Thailand and neighboring countries. *Mosq Systematics* **22**: 196–227.

Hewitt GM (2004). The structure of biodiversity—insights from molecular phylogeography. *Front Zool* **1**: 4.

Hope G, Kershaw AP, van der Kaars S, Xiangjun S, Liew P-M, Heusser LE *et al.* (2004). History of vegetation and habitat change in the Austral–Asian region. *Quatern Int* **118–119**: 103–126.

Irwin DE (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution* **56**: 2383–2394.

Iyengar A, Babu VN, Hedges S, Venkataraman AB, Maclean N, Morin PA (2005). Phylogeography, genetic structure, and diversity in the dhole (*Cuon alpinus*). *Mol Ecol* **14**: 2281–2297.

Knowles LL, Maddison WP (2002). Statistical phylogeography. *Mol Ecol* **11**: 2623–2635.

Kuhner MK (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**: 768–770.

Lu BL (1999). *Integrated Pest Management of Mosquitoes*. Science Press: Beijing.

Luo S-J, Kim J-H, Johnson WE, van der Walt J, Martenson J, Yuhki N *et al.* (2004). Phylogeography and genetic ancestry of tigers (*Panthera tigris*). *Plos Biol* **2**: e442.

Manguin S, Garros C, Dusfour I, Harbach RE, Coosemans M (2008). Bionomics, taxonomy, and distribution of the major malaria vector taxa of *Anopheles* subgenus *Cellia* in Southeast Asia: an updated review. *Infect Genet Evol* **8**: 489–503.

Moritz C, Patton JL, Schneider CJ, Smith TB (2000). Diversification of rainforest faunas: an integrated molecular approach. *Annu Rev Ecol Syst* **31**: 533–563.

O'Loughlin SM, Okabayashi T, Honda M, Kitazoe Y, Kishino H, Somboon P *et al.* (2008). Complex population history of two *Anopheles dirus* mosquito species in Southeast Asia suggests the influence of Pleistocene climate change rather than human-mediated effects. *J Evol Biol* **21**: 1555–1569.

Overgaard HJ, Ekbom B, Suwonkerd W, Takagi M (2003). Effect of landscape structure on anopheline mosquito density and diversity in northern Thailand: implications for malaria transmission and control. *Landscape Ecol* **18**: 605–619.

Overgaard HJ, Tsuda Y, Suwonkerd W, Takagi M (2002). Characteristics of *Anopheles minimus* (Diptera : Culicidae) larval habitats in northern Thailand. *Environ Entomol* **31**: 134–141.

Panchal M, Beaumont MA (2007). The automation and evaluation of nested clade phylogeographic analysis. *Evolution* **61**: 1466–1480.

Penny D (2001). A 40 000 year palynological record from northeast Thailand; implications for biogeography and palaeo-environmental reconstruction. *Palaeogeogr Palaeocl* **171**: 97–128.

Pfenninger M, Posada D (2002). Phylogeographic history of the land snail *Candidula unifasciata* (Poiret 1801) (Helicellinae, Stylommatophora): fragmentation, corridor migration and secondary contact. *Evolution* **56**: 1776–1788.

Powell JR, Caccone A, Amato GD, Yoon C (1986). Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. *Proc Natl Acad Sci USA* **83**: 9090–9093.

Pramual P, Kuvangkadilok C, Baimai V, Walton C (2005). Phylogeography of the black fly *Simulium tani* (Diptera: Simuliidae) from Thailand as inferred from mtDNA sequences. *Mol Ecol* **14**: 3989–4001.

Ray N, Adams JM (2001). A GIS-based vegetation map of the world at the last glacial maximum (25 000–15 000 BP). *Internet Archeology*. http://intarch.ac.uk/journal/issue11/.

Ray N, Currat M, Excoffier L (2003). Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* **20**: 76–86.

Rongnoparut P, Ugsang DM, Baimai V, Honda K, Sithiprasasna R (2005). Use of a remote sensing-based geographic information system in the characterizing spatial patterns for *Anopheles minimus* A and C breeding habitats in western Thailand. *Southeast Asian J Trop Med Public Health* **36**: 1145–1152.

Rosenberg NA, Nordborg M (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**: 380–390.

Sambrook J, Fritsch EF, Maniatis T (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor Press: New York.

Schneider S, Roessli D, Excoffier L (2000). *Arlequin ver. 2001: a Software for Population Genetics Data Analysis: Genetics and Biometry Laboratory*. University of Geneva: Switzerland.

Sharpe RG (1997). *The status of cryptic species within Anopheles minimus* PhD thesis University of Leeds, Leeds, UK.

Sharpe RG, Hims MM, Harbach RE, Butlin RK (1999). PCR-based methods for identification of species of the Anopheles minimus group: allele-specific amplification and single-strand conformation polymorphism. *Med Vet Entomol* **13**: 265–273.

Templeton AR (1998). Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol Ecol* **7**: 381–397.

Templeton AR (2008). Nested clade analysis: an extensively validated method for strong phylogeographic inference. *Mol Ecol* **17**: 1877–1880.

Templeton AR (2009). Why does a method that fails continue to be used? The answer. *Evolution* **63**: 807–812.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.

252

Van Bortel W, Trung HD, Sochantha T, Keokenchan K, Roelants P, Backeljau T *et al.* (2004). Eco-ethological heterogeneity of the members of the *Anopheles minimus* complex (Diptera: Culicidae) in Southeast Asia and its consequences for vector control. *J Med Entomol* **41**: 366–374.

Vanwambeke SO, Somboon P, Harbach RE, Ienstadt M, Lambin EF, Walton C *et al.* (2007). Landscape and land cover factors influence the presence of *Aedes* and *Anopheles* larvae. *J Med Entomol* **44**: 133–144.

Voris HK (2000). Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J Biogeog* **27**: 1153–1167.

Walton C, Sharpe RG, Pritchard SJ, Thelwell NJ, Butlin RK (1999). Molecular identification of mosquito species. *Biol J Linn Soc* **68**: 241–256.

Walton C, Handley JM, Tun-Lin W, Collins FH, Harbach RE, Baimai V *et al.* (2000). Population structure and population history of *Anopheles dirus* mosquitoes in southeast Asia. *Mol Biol Evol* **17**: 962–974.

Walton C, Somboon P, O'Loughlin SM, Zhang S, Harbach RE, Linton YM *et al.* (2007). Genetic diversity and molecular identification of mosquito species in the *Anopheles maculatus* group using the ITS2 region of rDNA. *Infect Genet Evol* **7**: 93–102.

White JC, Penny D, Kealhofer L, Maloney B (2004). Vegetation changes from the late Pleistocene through the Holocene from three areas of archaeological significance in Thailand. *Quatern Int* **113**: 111–132.

Zheng Z, Lei ZQ (1999). A 400 000 year record of vegetational and climatic changes from a volcanic basin, Leizhou Peninsula, southern China. *Palaeogeogr Palaeocl* **145**: 339–362.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)