Heredity (2008) 101, 186–195

www.nature.com/hdy

ORIGINAL ARTICLE

Forest-obligate *Sabethes* mosquitoes suggest palaeoecological perturbations

PM Pedro[1], MA Sallum[2] and RK Butlin[3]
[1]*Ecology and Evolution Group, School of Biology, University of Leeds, Leeds, UK;* [2]*Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brazil and* [3]*Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, UK*

The origin of tropical forest diversity has been hotly debated for decades. Although specific mechanisms vary, many such explanations propose some vicariance in the distribution of species during glacial cycles and several have been supported by genetic evidence in Neotropical taxa. However, no consensus exists with regard to the extent or time frame of the vicariance events. Here, we analyse the cytochrome oxidase II mitochondrial gene of 250 *Sabethes albiprivus* B mosquitoes sampled from western Sao Paulo in Brazil. There was very low population structuring among collection sites ($\Phi_{ST} = 0.03$, $P = 0.04$). Historic demographic analyses and the contemporary geographic distribution of genetic diversity suggest that the populations sampled are not at demographic equilibrium. Three distinct mitochondrial clades were observed in the samples, one of which differed significantly in its geographic distribution relative to the other two within a small sampling area ($\sim 70 \times 35$ km). This fact, supported by the inability of maximum likelihood analyses to achieve adequate fits to simple models for the population demography of the species, suggests a more complex history, possibly involving disjunct forest refugia. This hypothesis is supported by a genetic signal of recent population growth, which is expected if population sizes of this forest-obligate insect increased during the forest expansions that followed glacial periods. Although a time frame cannot be reliably inferred for the vicariance event leading to the three genetic clades, molecular clock estimates place this at ~ 1 Myr before present.
Heredity (2008) **101,** 186–195; doi:10.1038/hdy.2008.45; published online 28 May 2008

Keywords: cytochrome oxidase II; Pleistocene; Atlantic Forest; phylogeography; mtDNA; mosquitoes

Introduction

Tropical and subtropical forests are today experiencing change on a scale probably unseen since the end of the last Pleistocene glaciation. Although current concerns stem mainly from anthropogenic actions, the lessons learned from the past are highly relevant. The warming trends of the interglacials (including periods with temperatures higher than modern times), for example, may foreshadow the fate of forests exposed to the accelerated warming of recent decades. Similarly, the contentious hypotheses that forests were relegated to refugia or otherwise impacted during glacial periods (Haffer, 1969; Prance, 1982; Colinvaux, 1998; Colinvaux *et al*., 2000) are relevant to present-day habitat perturbations and may help to explain how ecosystems tolerate such changes. Moreover, most such theories imply that regions exist in the modern forest landscape that still hold high biodiversity because they remained ecologically stable throughout the Quaternary (Cracraft, 1985; Hall and Harvey, 2002).

It has been possible to use the genetic signatures from some organisms to track past demographic events including, notably, during the Pleistocene (Hewitt, 2000; Knowles, 2001; Lessa *et al*., 2003). These organisms are sufficiently sedentary that the patterns of genetic variation created by allopatry and subsequent range changes are not erased by subsequent gene flow. The Amazon basin has been the bread-and-butter for Neotropical phylogeography arguments, and very few data are available from other forest types, including the Atlantic Forest of coastal southern South America. Here we argue the utility of the forest-obligate *Sabethes albiprivus* B as one such model organism to track the history of this habitat.

Species of the genus *Sabethes* are diurnal New World culicines distributed from central Mexico to northern Argentina. All species are ecologically specialized to sylvan habitats (Forattini, 1965), a characteristic that makes them suitable for evaluations of changing forest distribution. They are exclusively phytotelmatic, using accumulated water in tree holes and bamboo internodes for oviposition. *Sa. albiprivus* B is restricted to the Atlantic Forest of southern Brazil.

Population demographic events leave a signature on DNA diversity that can be quantified by several means (Tajima, 1983; Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Griffiths and Tavare, 1994, 1996). One of these, the shape of the mismatch distribution, is often used as a guide to long-term demographic history (Rogers and Harpending, 1992). Stable populations will generally display ragged distributions resulting from the stochastic sampling that underwrites the drift-mutation process. In contrast, a population that has undergone recent demographic growth will display a unimodal

Correspondence: Dr PM Pedro, Faculdade de Saúde Pública, Universidade de São Paulo; Av. Dr Arnaldo, 715, São Paulo 01246-904, Brazil.
E-mail: pedrosquared@gmail.com
Received 13 December 2007; revised 26 March 2008; accepted 25 April 2008; published online 28 May 2008

distribution that persists until the population has remained stable for a long enough period of time. The mismatch distributions may be used in powerful tests of historical scenarios by comparing them with simulated data sets (Schneider *et al*., 2000).

We analysed genetic diversity of the cytochrome oxidase II (COII) mitochondrial locus in *Sa. albiprivus* B to infer contemporary and long-term demography using nested clade analysis (NCA) and maximum likelihood (ML) population estimates. We then used parametric bootstrapping to evaluate the plausibility of coalescent results. There was evidence of restricted dispersal of the species and a strong signature of long-term population fragmentation that preceded the last glacial maximum. We propose that *Sa. albiprivus* can provide an efficient way to identify areas that may once have been refuges and to track future habitat changes.

Materials and methods

Sample collection, sequencing and alignment

Sa. albiprivus B mosquitoes were collected during 2001 and 2002 in western Sao Paulo, Brazil (22° 35′ S, 52° 30′ W; Figure 1). This region is characterized by a matrix of interior Atlantic Forest patches surrounded by altered landscapes (primarily sugarcane and pasture) and bounded by two rivers, the Paraná in the northwest and Paranapanema in the south (Figure 1). Adult female mosquitoes were collected during daytime as they approached the collector seeking a blood meal. A summary of sampling locations is given in Table 1. Note that the suffix 'B' given in the species nomenclature is due to molecular evidence suggesting that museum voucher specimens of *Sa. albiprivus* differ from those sampled herein (Pedro, 2004).

The standard salting-out protocol with proteinase K digestion (Sambrook *et al*., 1989) was used to isolate total genomic DNA from whole, silica-desiccated mosquitoes. DNA was eluted in 20 μl of double-distilled water and a further dilution of 1:40 yielded the working stock concentration. The universal insect primer pair TK-N-3785 (5′-GTTTAAGAGACCAGTACTTG-3′) and TL2-J-3037 (5′-ATGGCAGATTAGTGCAATGG-3′) (Simon *et al*., 1994) was used to amplify a 748-bp region of the mitochondrial DNA (mtDNA) including the entire *COII* gene from 250 *Sa. albiprivus* B individuals. This was undertaken in a reaction volume of 25 μl using DNA from approximately 1/100th of a mosquito. Final PCR concentrations were: 0.2 μM deoxyribonucleotide triphosphates, 0.6 μM of each primer, 1.5 mM $MgCl_2$ and one unit of Taq. Cycling conditions included an initial denaturation at 94 °C for 4 min; 35 cycles of denaturation at 94 °C for 1 min, annealing at 52 °C for 1 min and extension at 72 °C for 1 min and a final extension step at 72 °C for 4 min. The PCR products were cleaned using standard columns (Promega, Madison, WI, USA) and sequenced in one direction using primer TK-N-3785 on an ABI 377 automated sequencer (PE Applied Biosystems, Warrington, England). The 447 nucleotide positions unambiguously readable for all sequences were aligned using the CLUSTALW (Thompson *et al*., 1994) peripheral of the Bioedit version 7.00 suite of programs (Hall, 1999). Ambiguous sequences and the 10 singleton haplotypes with non-synonymous substitutions were re-sequenced to confirm initial chromatograph readings. The presence of pseudogenes in the data set was discounted because no stop codons were identified. The sequences for 61 unique mtDNA haplotypes are available on the public databases (GenBank accession no. EU563021–EU563081).

Statistical analyses

We used Modeltest version 3.06 (Posada and Crandall, 1998) with a model block implemented in PAUP 4.0b (Swofford, 1993) to find the most suitable model of COII evolution. The TrN (Tamura and Nei, 1993) model was selected, with the following estimated parameters: substitution rate matrix of A↔C:1.00, A↔G:35.09, A↔T:1.00, C↔G:1.00, C↔T:14.08 and G↔T:1.00; γ-shape parameter of 0.109 and no invariable sites.

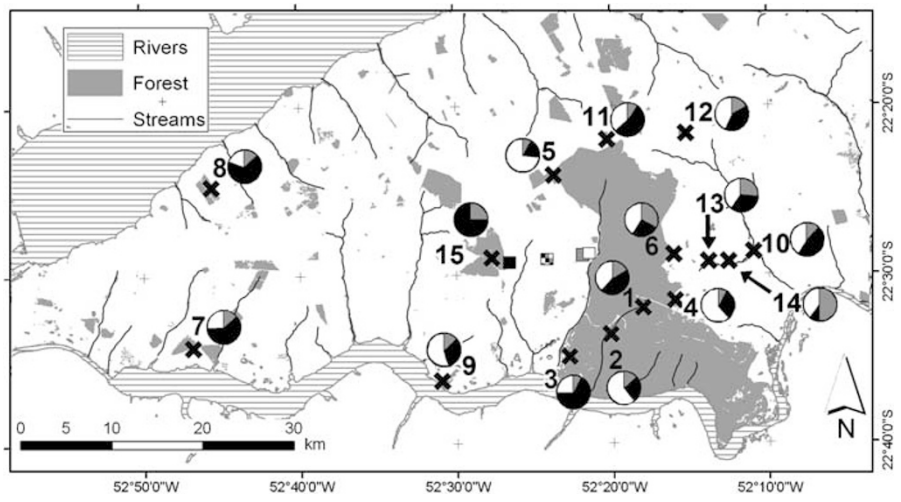

Figure 1 Map of collection sites for all *Sa. albiprivus* B sampled in western Sao Paulo. Pie charts represent each of the three main mitochondrial DNA (mtDNA) clades (as shown in Figure 2): black, 3−1; grey, 3−2 and white, 3−3. The geographic centre, calculated according to Templeton *et al*. (1995), of each clade is represented by an appropriately coloured square and the total geographic centre is the multi-coloured square. Horizontal length of image is approximately 100 km.

Table 1 Collection sites, coordinates, number of samples, haplotype composition, area of sampled forests and genetic diversity estimates (for $N>9$)

Sampling location	*Coordinates*	N	*Haplotypes*	*Approximate forest size (m^2)*	θ_S	*Gene diversity*
1	22 31 33 S, 52 18 17 W	35	A01$^{(5)}$,A03$^{(1)}$,B01$^{(8)}$,B02$^{(3)}$,B05$^{(1)}$, B07$^{(1)}$,**B10**$^{(1)}$,**B17**$^{(1)}$, **B18**$^{(1)}$,C01$^{(5)}$, C02$^{(2)}$,**C12**$^{(1)}$,**C13**$^{(1)}$, **C14**$^{(1)}$,**C18**$^{(1)}$,**C21**$^{(1)}$,**C22**$^{(1)}$	400 000 000	6.19 ± 3.35	0.91 ± 0.03
2	22 32 29 S, 52 19 19 W	15	A01$^{(1)}$,A04$^{(1)}$,B01$^{(3)}$,B02$^{(1)}$,C01$^{(2)}$, C02$^{(4)}$,C08$^{(1)}$,**C25**$^{(1)}$,**C30**$^{(1)}$	400 000 000	6.50 ± 3.62	0.90 ± 0.05
3	22 33 17 S, 52 22 11 W	15	**A08**$^{(1)}$,B01$^{(5)}$,**B03**$^{(2)}$,B07$^{(1)}$,**B08**$^{(1)}$, **B14**$^{(1)}$, C01$^{(1)}$,C02$^{(2)}$, **C32**$^{(1)}$	400 000 000	6.40 ± 3.51	0.89 ± 0.07
4	22 31 04 S, 52 16 24 W	15	A01$^{(1)}$,B01$^{(3)}$,B02$^{(2)}$,C01$^{(2)}$,C02$^{(1)}$, C03$^{(1)}$,C04$^{(1)}$,C05$^{(1)}$,**C06**$^{(2)}$,**C27**$^{(1)}$	400 000 000	6.03 ± 3.29	0.94 ± 0.04
5	22 24 39 S, 52 22 59 W	15	A01$^{(1)}$,B01$^{(1)}$,B02$^{(1)}$,**B13**$^{(1)}$,C01$^{(3)}$, C02$^{(5)}$,C03$^{(1)}$,**C10**$^{(1)}$,**C11**$^{(1)}$	400 000 000	5.90 ± 3.35	0.88 ± 0.07
6	22 28 51 S, 52 16 11 W	15	A01$^{(3)}$,A02$^{(1)}$,A04$^{(1)}$,B01$^{(3)}$,B02$^{(1)}$, C01$^{(2)}$,C02$^{(1)}$,C03$^{(1)}$,C07$^{(1)}$,**C17**$^{(1)}$	400 000 000	5.71 ± 3.25	0.93 ± 0.05
7	22 34 00 S, 52 46 50 W	30	A01$^{(2)}$,**A05**$^{(2)}$,B01$^{(12)}$,B02$^{(4)}$,B4$^{(1)}$, **B15**$^{(1)}$,C01$^{(2)}$,C02$^{(1)}$,C3$^{(2)}$, C7$^{(1)}$, **C16**$^{(1)}$,**C31**$^{(1)}$	4 600 000	5.78 ± 3.38	0.83 ± 0.06
8	22 24 25 S, 52 46 10 W	15	A01$^{(2)}$,B01$^{(5)}$,B02$^{(2)}$,**B06**$^{(2)}$,**B11**$^{(1)}$, C01$^{(2)}$,C08$^{(1)}$	3 920 000	5.39 ± 3.09	0.87 ± 0.06
9	22 35 59 S, 52 30 16 W	15	A01$^{(2)}$,B01$^{(3)}$,B02$^{(1)}$,**B19**$^{(1)}$,C01$^{(4)}$, C02$^{(1)}$,**C26**$^{(1)}$,**C28**$^{(1)}$,**C29**$^{(1)}$	1 150 000	6.08 ± 3.44	0.90 ± 0.05
10	22 29 07 S, 52 11 23 W	17	A01$^{(1)}$,**A06**$^{(1)}$,B01$^{(7)}$,B02$^{(1)}$,C01$^{(4)}$, C03$^{(1)}$,C09$^{(1)}$,**C33**$^{(1)}$	440 000	5.39 ± 3.09	0.80 ± 0.08
11	22 22 08 S, 52 20 18 W	11	A01$^{(1)}$,B01$^{(5)}$,B02$^{(1)}$,C01$^{(1)}$,C04$^{(1)}$, C09$^{(1)}$,**C34**$^{(1)}$	225 000	4.88 ± 2.82	0.82 ± 0.12
12	22 21 40 S, 52 15 34 W	23	A01$^{(1)}$,A02$^{(2)}$,**A07**$^{(1)}$,B01$^{(7)}$,B02$^{(1)}$, **B09**$^{(1)}$,C01$^{(3)}$,C02$^{(4)}$,**C19**$^{(1)}$,**C20**$^{(1)}$,**C24**$^{(1)}$	150 000	6.40 ± 3.75	0.88 ± 0.05
13	22 29 30 S, 52 13 35 W	15	A01$^{(3)}$,A03$^{(1)}$,B01$^{(3)}$,B02$^{(1)}$,**B16**$^{(1)}$, C01$^{(2)}$,C02$^{(3)}$,C03$^{(1)}$	90 000	6.67 ± 3.74	0.90 ± 0.05
14	22 29 16 S, 52 12 56 W	10	A01$^{(5)}$,**B12**$^{(1)}$,C02$^{(1)}$,C05$^{(1)}$,**C15**$^{(1)}$, **C23**$^{(1)}$	40 000	6.95 ± 3.88	0.78 ± 0.14
15	22 29 25 S, 52 27 36 W	4	A01$^{(1)}$,B02$^{(1)}$,B04$^{(1)}$,B05$^{(1)}$	NA	NA	NA

Abbreviation: NA, not available.
Underlined haplotypes were found in only one sampling location (private alleles).
Haplotype superscripts in parentheses indicate total number sampled.

Nucleotide frequencies were estimated to be A:0.40, C:0.11, G:0.15 and T:0.34. Unless otherwise noted, these ML parameters were used for all downstream analyses.

Arlequin version 2.001 (Schneider *et al*., 2000) was used to calculate the mitochondrial haplotype diversity index h (Nei, 1987) and nucleotide diversity θ_S. This program was also used to test for deviations from neutrality in the COII locus using Tajima's D and Fu's F_S tests. The R^2 statistic (Ramos-Onsins and Rozas, 2002) was calculated using the program DNASP (Rozas *et al*., 2003), with significance based on the null hypothesis of a neutral population estimated through 1000 coalescent simulations. An F_{ST}-based analysis of molecular variance (AMOVA) was implemented in ARLEQUIN using the 15 sampled subpopulations to evaluate contemporary population structuring. An analogous AMOVA using the Φ_{ST} measure of population differentiation, which weights sequence similarity (Reynolds *et al*., 1983), was also undertaken using the ML parameters described above. The program ZT (Bonnet and Peer, 2002) was used in Mantel tests of association between the log-scale geographic distance and either $F_{ST}/(1-F_{ST})$ or $\Phi_{ST}/(1-\Phi_{ST})$ (Rousset, 2000). Here, P-values were estimated by comparison with 10 000 random permutations.

We did not perform phylogenetic analyses on the COII data because intraspecific lineages are not bifurcating and such analyses do not normally consider ancestral haplotypes as still extant (Posada and Crandall, 2001). Instead we used the statistical parsimony algorithm (Templeton, 1998) in TCS version 1.13 (Clement *et al*., 2000). Subsequent to this, NCA was used to test for a non-random geographical distribution of haplotypes (Templeton, 1998; Templeton *et al*., 1995). This was implemented in Geodis version 2.0 (Posada *et al*., 2000) with significance evaluated by 10 000 random permutations of clades among sampling sites. NCA estimates clade distance (D_c), which measures the geographical spread of a clade, and the nested clade distance (D_n), which measures how a clade is geographically distributed relative to others in the same nesting level (that is temporally related nodes). The analysis also compares the average distance of all tip clades to those of the internal clades within a nesting level (I-T), a statistic that, when significantly positive, indicates that younger tip clades are less geographically dispersed than older interior clades (providing evidence for restricted gene flow).

Homoplasic loops in the TCS network were corrected in one of two ways. We first re-calculated connections using nucleotide positions beyond 447 bp from several unambiguous sequence traces. Otherwise, loops were resolved such that the haplotype having the highest outgroup weight in each node (based on TCS output) was considered ancestral and thus internal in the

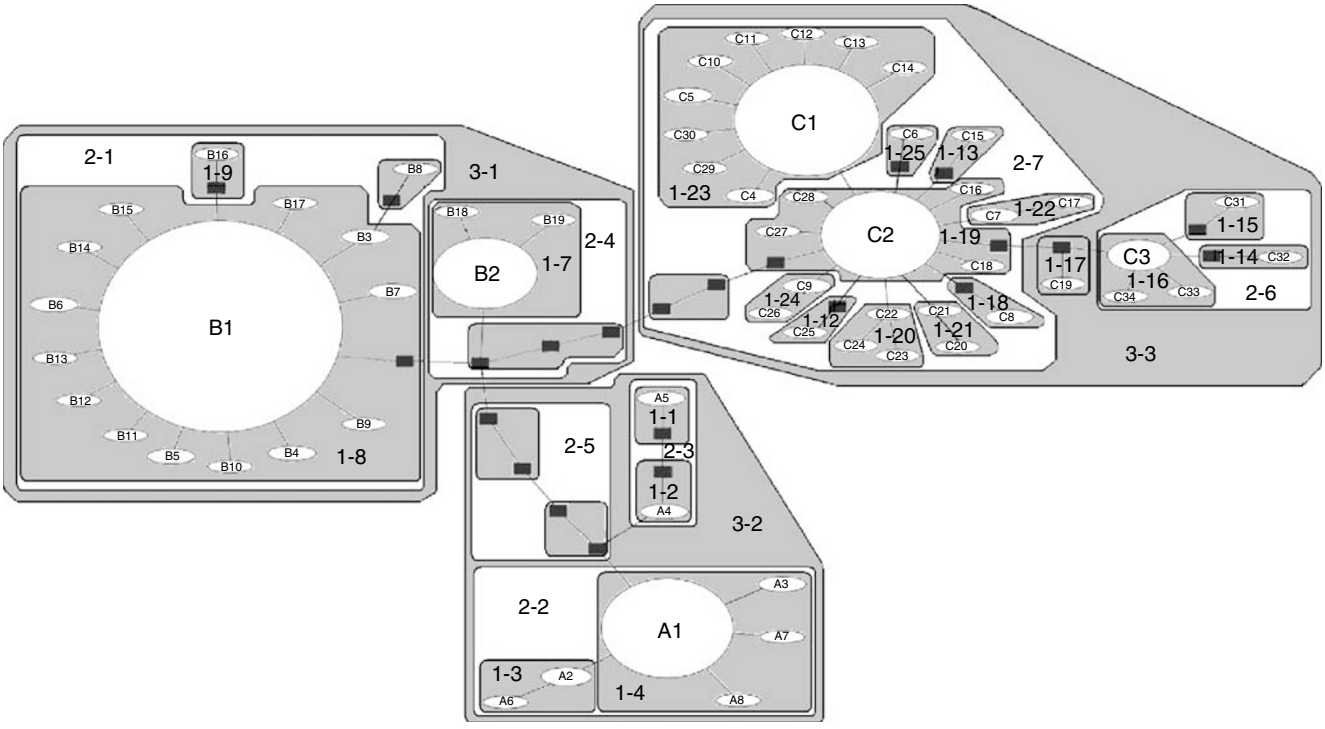

Figure 2 TCS minimum spanning network following correction for homoplasy. Hierarchical groupings of clades for nested clade analysis (NCA) are also shown.

network, as proposed by Castelloe and Templeton (1994). Due to the presence of very deep inter-clade branches, we estimated outgroup probabilities separately for each haplogroup (that is haplogroups A, B and C in Figure 2).

The net average divergence (D_a) (Nei and Li, 1979) among *Sa. albiprivus* B mtDNA clades was estimated using Mega version 2.1 (Kumar *et al*., 2001). Divergence time among haplogroups was calculated as D_a divided by 2.3%, the generic arthropod mitochondrial divergence rate per million years (Brower, 1994).

Simultaneous estimates of mtDNA population size (calculated from the ML estimate of θ) and growth rate were derived using Fluctuate version 1.4 (Kuhner *et al*., 1998). The parameter θ represents $2N_{ef}\mu$ for mtDNA, where N_{ef} is the effective female population size and μ is the sequence mean substitution rate over the 447 bp of the *COII* gene used. The population-growth parameter, g, describes the direction and magnitude of population-size change over time measured in units of $1/\mu$ per generation. The ML values for the *Sa. albiprivus* B data were estimated under the HKY85 (Hasegawa *et al*., 1985) model in PAUP* (data not shown) and the initial θ value was calculated based on Watterson (1975). The permutations were run for 20 short (20 000 steps, sampling every 100 generations) and 2 long chains (250 000 steps, sampling every 250 generations) to estimate g and θ in four independent runs to confirm convergence.

Although Fluctuate finds the best population parameters (θ and g) to explain the observed gene topology, its output is limited by the models that it can simulate: either stable or exponential change. We thus used Simcoal version 1 (Excoffier *et al*., 2000) in a parametric bootstrap analysis to test whether the observed data are a good fit to the two elementary models. One-thousand data sets were created to quantify the probability that a simulated population could arrive at the observed genealogical topology under each of the two hypotheses. The transition:transversion ratio and the shape of the γ-distribution were estimated by ML (10.35 and 0.072, respectively) in PAUP* and base frequencies were maintained at 0.25. The simulated data sets represented the coalescent expectations based on Fluctuate output parameters and were compared to the observed *Sa. albiprivus* B data using the mismatch distribution. We used the sum of squared differences (SSD) as a test statistic (see ARLEQUIN manual; Schneider *et al*., 2000). These analyses were performed with GenStat 6 (VSN International Ltd).

Results

Genetic diversity and population structure

The fifteen *Sa. albiprivus* B collections displayed little inter-population variation in standard summary statistics. DNA haplotype and nucleotide diversity was high in all forest fragments surveyed (Table 1). Although there was a tendency for gene diversity to be lower in the smaller fragments, there was no significant association between forest size and diversity indices ($r_s = 0.40$, $P > 0.10$). The frequency-based AMOVA indicates that 99% of all the variation in the 15 samples is present within populations, suggesting a scenario close to panmixia ($F_{ST} = 0.01$, $P = 0.11$). An analysis that included weighting of sequence differentiation revealed weak, but significant, population structure ($\Phi_{ST} = 0.03$, $P = 0.04$). An evaluation of mtDNA structure within the largest forest fragment alone (see Figure 1) did not yield significant AMOVA values for either distance measure, $F_{ST} = -0.002$ ($P = 0.57$) or $\Phi_{ST} = 0.04$ ($P = 0.07$). This is

Table 2 Inference chain for the significant geographical associations of clades from NCA analysis

Clade	*Chain of inference*	*Inference*
Haplotypes nested in 1–4	1,2,11,17 (no)	Inconclusive outcome
Haplotypes nested in 1–8	1,2,3,5,6,7 (yes)	Restricted gene flow/dispersal but with some long-distance dispersal
Haplotypes nested in 1–19	1,2,11,17 (no)	Inconclusive outcome
Haplotypes nested in 1–23	1,2,3,4 (no)	Restricted gene flow with isolation by distance
Clades in entire cladogram	1,2,3,4 (no)	Restricted gene flow with isolation by distance

See reference Templeton (2004).

potentially a result of lower sample size rather than difference in structure between the subdivided and continuous habitats.

Mantel tests for the association of genetic and geographic distance using all sampled forest fragments were not significant using either F_{ST} ($r = -0.02$; $P = 0.55$) or Φ_{ST} ($r = 0.07$; $P = 0.35$). Tests of IBD including only data for the six sampled locations within the largest forest revealed a significant association with geographic distance using F_{ST} ($r = 0.56$; $P < 0.05$), but no isolation by distance (IBD) was observed with Φ_{ST}. Our permutation test for the localization of singleton/doubleton, non-ancestral haplotypes showed that these recently derived haplotypes are located closer together within the entire collection site than is expected by chance ($P < 0.01$).

Phylogeographic inferences

The TCS output yielded a set of six common haplotypes surrounded by less-common satellite haplotypes. The latter included fourteen doubletons and one tripleton (haplotypes A3, A4, A5, B3, B4, B5, B6, B7, C4, C5, C6, C7, C8, C9 and A2 in Figure 2). Such starburst patterns have been associated with scenarios of recent population expansion (Rogers and Harpending, 1992). The depth of the branches connecting each of the six common haplotypes led to our subclassification of the network into three haplogroups: A, B and C. Inter-haplogroup average pairwise distances were 7.24 between haplogroups A and B, 8.98 between B and C and 10.30 between A and C. The mean intra-haplogroup pairwise distances were: 1.06 for A, 1.42 for B and 2.06 for C. These three groupings also corresponded to the three third-nesting-level clades in the NCA (see below). The estimated divergence estimated in Mega under incorporating the TrN evolutionary model within Mega ($\pm$ s.e.) between haplogroups A and B was 0.019 ± 0.01 (0.95×10^6 years), between A and C 0.029 ± 0.013 (1.45×10^6 years) and between B and C 0.024 ± 0.012 (1.2×10^6 years).

NCA was used to test the interaction between phylogenetic age and geographic location of haplotypes. Three out of five significant clade distances suggested restricted gene flow leading to isolation-by-distance because younger (tip) clades are significantly less widely distributed than ancestral (interior) clades (Table 2; Supplementary Table 1). Haplotypes within clade 1–8 appear to show restricted gene flow as the I-T distance based on both D_C and D_N is significantly large, which suggests that the internal haplotype (B1) is significantly more distributed than its satellite clades. Similarly, clade 1–23 indicates restricted gene flow because of the significantly wider distribution of the internal C1 relative to the others in that nesting clade. The highest-order clade (full cladogram) also suggests restricted gene flow with IBD. It indicates a significantly more distributed interior clade (3−1) than either of the other two third-order clades, which is significantly large D_C. Furthermore, 3−1 also has significantly large D_N, meaning that its haplotypes are not only more widely distributed from its own geographic centre, but also that its centre is distinct from the geographic centre of all haplotypes considered. Clade 3−3 shows significantly small D_C, indicating that it is more restricted in distribution than either 3−1 or 3−2. The I-T, D_C and D_N distances also support the relatively wider distribution of clade 3−1 (but see 'Discussion' for a caveat with regard to this inference). Using the algorithm of Templeton *et al*. (1995), we calculated the centres of each third-order clade and found that, for clade 3−1, this was approximately 9 km east of either clade 3−2 or 3−3 (which were nearly coincident) (Figure 1).

Although NCA highlights significant associations within a network, the method explicitly assumes independence of nesting clades, such that 'all distances and statistical assessments are performed within nested groups of evolutionarily close clades' (Templeton *et al*., 1995). In light of this limitation, we assessed whether all recently derived haplotypes generally display a pattern of geographic localization, as would be expected if new haplotypes were produced in the population faster than they are propagated by gene flow (that is isolation-by-distance). We considered the geographic distribution of only the fourteen doubleton and one tripleton haplotypes, as these tip nodes are expected to represent the most novel mutations. We averaged the geographic distances between collections of these haplotypes and compared this to 1000 randomly selected combinations of the same number of haplotype groups drawn from the 15 sample sites in proportion to sample sizes to test the null hypothesis of panmixia. This analysis detected a significant association between the derived haplotypes and geographic proximity ($P < 0.01$), supporting (for the full cladogram) the occurrence of local IBD inferred by the NCA (Table 2).

Demographic estimates

The Fluctuate ML estimate of female effective population size under a stable population model was 23 million ($\theta = 0.046$ (95% CI: 0.035 and 0.053)). Assuming an exponentially growing population model, the estimated growth rate (r) for the *Sa. albiprivus* B data was 3.87×10^{-7} (s.d. $= 3.15 \times 10^{-8}$) and current female effective population size was 37.5×10^6 (s.d. $= 1.6 \times 10^6$). This represents a sample from a population undergoing only slight population increase.

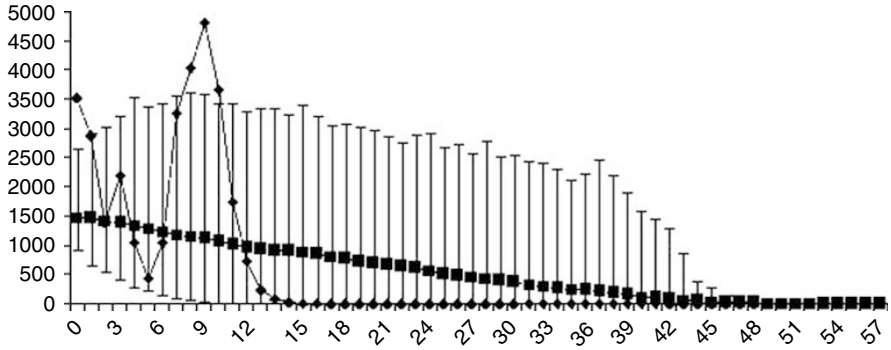

Figure 3 Distribution of the mean of 1000 simulated samples at constant population size (black squares) plotted with the 95th percentile values and with the observed distribution superimposed (black diamonds). The *x* axis indicates number of mismatches and *y* axis their frequencies.

The measures of fit to the standard neutral expectations for COII proved discordant. Tajima's *D*-value of −0.80 and the R^2 value of 0.058 were not significant ($P = 0.22$ and 0.21, respectively) and, therefore, consistent with neutrality and a stable population size. However, Fu's F_S of −24.72 proved highly significant ($P < 0.001$).

Demographic history was also estimated separately for haplogroups A, B and C because of the possibility that these did not exist as an interbreeding population as their MRCA. Here, estimates point to large, growing populations in all cases. Approximate Fluctuate population sizes for clades A, B and C are 1×10^7, 1.1×10^9 and 2.3×10^9, respectively. The growth constants for each haplogroup were 9.9×10^{-7} for A, 1.7×10^{-6} for B and 9.5×10^{-7} for C. Tajima's *D*, Fu's F_S and R^2 tests also suggest recent population growth, with all simulations proving significant except Tajima's *D* for haplogroup A.

Testing putative population histories

Although coalescence times in structured populations tend to increase relative to unstructured populations (Avise, 2000), AMOVA analyses herein suggest that current genetic structure is a not a major factor in the *Sa. albiprivus* B population sampled. We thus conducted our analyses treating the entire field site as a single interbreeding population. Our first hypothesis tested the fit of the observed *Sa. albiprivus* B data to a modelled population of constant size (that is one in which the population has remained at mutation-drift equilibrium since its most recent common ancestor (MRCA)). We used Simcoal to generate 1000 coalescent simulations of a population of effective size 23×10^6 (the Fluctuate output from above). As expected theoretically (Rogers and Harpending, 1992), the simulated equilibrium populations resulted, on average, in an approximately L-shaped mismatch distribution (Figure 3). The simulated distributions show marked deviations from the observed frequencies and contain substantially more mismatch categories, on average, than the *Sa. albiprivus* B data, which has 18. The average of the observed SSDs was significantly different from that of the simulated values ($P = 0.007$). We therefore rejected the possibility that the *Sa. albiprivus* B sequences were sampled from a stable population of 23 million that had been so since its MRCA.

We next undertook 1000 coalescent Simcoal simulations based on the Fluctuate ML estimates for a growing population. The average of the resulting mismatch distributions was, as expected theoretically, unimodal (Figure 4). The average SSD value for the observed data was significantly different from that of the simulated data set ($P = 0.012$), leading to the rejection of the hypothesis that the mosquitoes sampled were from a population undergoing exponential growth since MRCA.

We tested the sensitivity of our conclusion to uncertainty in the estimates of population size and growth rate by running additional Simcoal simulations with population sizes $100N$, $10N$, $2N$, N, 0.5, $0.1N$ and $0.01N$ and growth rates 100, 10, 2, 0.5, 0.1 and $0.01\,g$, where N and g are the estimates obtained from Fluctuate. In all cases, the observed distribution of mismatches differed from the simulated distribution ($P < 0.01$).

Discussion

Results from the Simcoal parametric simulations suggest that a simple demographic history is unlikely to explain the frequencies and phylogenetic arrangement of COII haplotypes in the population studied. The feature of the data responsible for the lack of fit to either a stable or exponentially growing population is likely the very deep branches among haplogroups A, B and C.

Arguably, the three haplogroups may each represent distinct species. We tested this possibility by sequencing an intron of the (nuclear) actin gene for several representatives from each of the three haplogroups. There was no significant variation in nuclear sequences among groups defined by the mtDNA haplotypes, and thus we failed to reject the null hypothesis of a single species (Pedro, 2004). Moreover, similar intraspecific distances have been observed in the COII marker for other species of mosquito (Chen *et al*., 2003), so the extent of the inter-clade divergence is not aberrant.

Evidence for allopatry

Inferences with regard to the demographic history of the sampled *Sa. albiprivus* B subpopulations must be evaluated in light of the significant results seen in the highest-level clades using NCA (Table 2). Particularly relevant to an argument of past allopatry is the fact that the geographic distribution of the three major haplogroups is not random. Although NCA indicates this

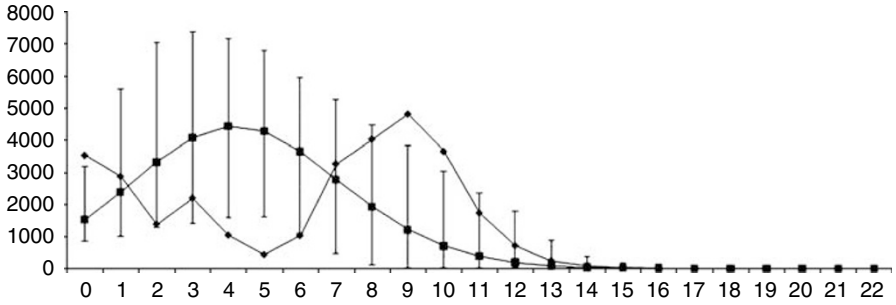

Figure 4 Distribution of the mean of 1000 simulated populations of exponentially growing size (black squares) plotted with the 95th percentile values and with the observed distribution superimposed (black diamonds). The *x* axis indicates number of mismatches and *y* axis their frequencies.

pattern to be IBD, the reliability of the inference key has been questioned (Knowles and Maddison, 2002; Masta *et al*., 2003; Panchal and Beaumont, 2007; Petit, 2008) and it has undergone alterations since its introduction, as alternative phylogeographic scenarios have been proposed (Templeton, 2004). A diagnosis of IBD at the highest nested level may not be reliable for at least three reasons as follows.

Although IBD may be expected, as the lowest-level clades showed a similar signal (Table 2), there is a lack of IBD in the intermediate levels of the network (nesting levels 1 and 2). Templeton *et al*. (1995) suggest that true IBD will normally be seen in the lowest levels and will either be absent or gradually be attenuated with increasing nesting levels. Intuitively, this means that as one moves to higher nesting levels, looking farther into the past, the signal of IBD will gradually be lost because old haplogroups are expected to be less geographically localized than the younger ones.

The inference of IBD in Templeton's key is based on the coalescent expectation that older haplotypes/clades (those positioned internally in the network) will be more broadly distributed than younger, more derived ones (external in the network structure). However, evaluation of what is derived and what is ancestral at the level of the entire network (Figure 2) is unreliable in this case because of the deep internal branches among the major haplogroups. Indeed, all clades are most likely derived from an ancestral node that is no longer extant.

If the IBD inferred at the level of the full haplotype network were indeed correct, it should also be apparent in the Mantel test using Φ_{ST} (which incorporates information on haplotype similarity) among all subpopulations, which it is not.

Thus, at the lowest nesting level, restricted gene flow perhaps indeed results from IBD (entries for nesting clade 1–8 and 1–23 in Table 2). This is supported by our analysis of the geographical localization of young haplotypes and by the significant correlation between F_{ST} and geographic distance within the largest forest fragment. However, at the highest nesting level, we suggest that, instead of IBD, the contemporary signature in the sampled haplotypes is most readily explained by historical allopatry of three higher-order clades (haplogroups A, B and C) that only recently came into secondary contact: although their distributions now overlap extensively, they retain a signature of their independent expansion into the sampled area. Templeton's key does not detect this past fragmentation because it requires that sampling areas exist where the clades do not overlap (Templeton, 2004). Thus, although the GeoDis inference key is prone to limitations, the program nevertheless generates statistics that can be useful in evaluating the significance of patterns of the geographical distribution of haplotypes.

Following expansion and secondary contact, diffusion of one haplotype into the range of another is expected to be slow under the phalanx model of dispersal (Endler, 1977; Ibrahim *et al*., 1996). It is thus unsurprising that haplogroups have not had time to fully permeate the field site and thus that some clades are less represented than others in certain areas (as indicated by NCA). The offset distributions of each clade are apparent in Figure 1 where haplogroup B (nesting clade 3−1) is relatively more westerly distributed than the other two (3−2 and 3−3). This suggests that at least one population (B haplotypes) expanded its range from a western area into the present-day field site. Clade 3−1 is also significantly more widely distributed in the field site (Supplementary Table 1) than the other haplogroups, possibly because it was sampled closer to the original site of its refugium. Likewise, Clade 3−3, which is least widely distributed geographically, may have had a refugium farther from the sampled sites.

Our additional analysis using only derived doubletons and one tripleton also suggests that the dispersal capacity of *Sa. albiprivus* B is low as pairs of identical haplotypes are more geographically proximate than expected by chance. Likewise, the IBD found within the largest forest (using Mantel tests) suggests that the neighbourhood (deme) size for female *Sa. albiprivus* B is relatively small (that is smaller than the size of this largest fragment), perhaps a surprising result for a flying insect. The lack of significant IBD across the whole sample area may be because this scale was too large relative to dispersal distance or because forest fragmentation has interrupted gene flow. It is, however, unlikely to be related to genetic drift within small fragments because diversity was not related to forest fragment size for this forest-obligate species.

It is feasible that the three main haplogroups are derived from the bottlenecking of a single large population, a scenario that would have removed all but the three divergent haplogroups. However, this cannot account for the significantly different geographic distribution of each haplogroup.

The starburst patterns seen throughout the network (Figure 2), where a central common haplotype is surrounded by several less frequent satellite haplotypes, are consistent with a rapidly expanding population (Avise, 2000). Ramos-Onsins and Rozas (2002) noted that simulations using large sample sizes ($\geqslant 50$) yield significantly negative F_S values for neutral loci in growing populations, as was observed for *Sa. albiprivus* B. Thus, it is probable that the observed data set encompasses samples from a growing population, as the majority of mtDNA variation is neutral (but see William *et al*., 1995; Hillis *et al*., 1996).

The palaeoecological framework

Evidence for rapid population expansions during the late Pleistocene has been found in several species (Good and Sullivan, 2001; Hundertmark *et al*., 2002; Zheng *et al*., 2003). Most, however, are seen in temperate latitudes and presumed to result from the attenuation of glaciation and its direct consequences, such as ice cover and permafrost. A demographic effect has rarely been explicitly identified in regions that experienced only secondary consequences of the ice ages, such as Neotropical forests, where there was likely substantial ecological turnover, but no glaciation. However, genetic evidence exists from other forest-specialized mosquito species consistent with a late Pleistocene expansion, although the separation between divergent haplotypes proceeds the last ice age (Mirabello and Conn, 2006; Conn and Mirabello, 2007). Moreover, these examples generally implicate the location of refugia in Amazonian Brazil rather than the Atlantic Forest.

There are at least two consequences of the allopatric refuge hypothesis for species specialized to humid forests. First, there should exist centres of genetic endemism (Vane-Wright *et al*., 1991; Moritz and Faith, 1998; Evans *et al*., 2003) surrounded by areas of attenuated diversity. Samples from adjacent centres of endemism should show levels of genetic divergence consistent with their time in allopatry. Second, forest-obligate organisms should display the genetic imprint of a recent population expansion (following habitat amelioration). The first expectation has rarely been tested, but the few Neotropical data sets that exist (Glor *et al*., 2001; Mallarino *et al*., 2005) usually suggest that closely related sister species diverged substantially earlier then the last glacial maximum (in agreement with the time frame observed here for *Sabethes*). A recent analysis by Lessa *et al*. (2003) identified equivocal genetic signals from western Amazonian mammals, most of which did not show evidence of recent population growth (as would be expected if favourable habitat increased with warming conditions). To our knowledge, our data represent the first evidence of significant population increase and potential allopatry during the Pleistocene for an Atlantic Forest species.

If the deep branches seen among the three *Sa. albiprivus* B clades result from allopatry, sequence divergence suggests an approximate time frame of 0.95, 1.2 and 1.45 Myr between A and B, B and C and A and C, respectively. Thus in each of these cases the clades would have been isolated for a substantial portion of the entire Pleistocene, a period of time that would include several ice ages. It is thus possible that the historic population allopatry inferred herein may result from the superimposition of several cycles of habitat contraction and re-expansion.

Conservation implications

Mosquitoes provide a potentially robust appraisal of habitat health because they have direct and indirect links to many ecological niches. Female mosquitoes require blood meals from vertebrates (primates and birds in the case of *Sabethes*) to complete oogenesis. Furthermore, both sexes obligately feed on flower and fruit nectars. *Sabethes* provide a further, indirect, measure of forest age and regeneration level as several species require tree holes bearing accumulated water for oviposition, such that they will be scarce in secondary forests lacking trees of sufficient size.

Although fragmented, or otherwise altered, forest habitats were no doubt different things to different organisms, centres of endemism should today still maintain high inter- and intra-specific diversity. This is because they have provided a stable haven through various cycles of forest constriction/alteration and re-expansion. Moreover, species diversity is expected to be highest in areas of the historical refugia because it takes time for taxa to disperse into newly forested regions following habitat expansions and to accumulate high levels of genetic diversity (Ibrahim *et al*., 1996). These centres of endemism should therefore be considered conservation priorities. Areas of overlap that have received immigrants from multiple refugia, as we posit here, should also possess substantial diversity because of the admixture of divergent haplogroups.

Acknowledgements

We are indebted to funding assistance from Sigma Xi and the Food, Drug and Health Foundation. PMP is grateful for a PhD studentship from the University of Leeds. Ralph E Harbach at The Natural History Museum, London, provided considerable contributions to the morphological identification of animals. We also thank Cathy Walton for much input and discussion. Also acknowledged for their contribution are Mariana Vale, Alexandre Uezu and two anonymous reviewers. Sampling was undertaken with a Brazilian Federal Permit (Processo número 205/2001—CGFAU/LIC Processo número 02001.007995/00-65).

References

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

Bonnet E, Peer YVd (2002). zt: A software tool for simple and partial Mantel tests. *J Stat Software* **7**: 1–12.

Brower AVZ (1994). Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial-DNA evolution. *Proc Natl Acad Sci USA* **91**: 6491–6495.

Castelloe J, Templeton AR (1994). Root probabilities for intraspecific gene trees under neutral coalescent theory. *Mol Phylogenet Evol* **3**: 102–113.

Chen B, Butlin RK, Harbach RE (2003). Molecular phylogenetics of the oriental members of the Myzomyia series of *Anopheles* subgenus *Cellia* (Diptera: Culicidae) inferred from nuclear and mitochondrial DNA sequences. *Syst Entomol* **28**: 57–69.

Clement M, Posada D, Crandall KA (2000). TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**: 1657–1659.
Colinvaux PA (1998). A new vicariance model for Amazonian endemics. *Global Ecol Biogeogr* **7**: 95–96.
Colinvaux PA, De Oliveira PE, Bush MB (2000). Amazonian and neotropical plant communities on glacial timescales: the failure of the aridity and refuge hypotheses. *Quaternary Sci Rev* **19**: 141–169.
Conn JE, Mirabello L (2007). The biogeography and population genetics of neotropical vector species. *Heredity* **99**: 245–256.
Cracraft J (1985). Historical biogeography and patterns of differentiation within the South American areas of endemism. *Ornithol Monogr* **36**: 49–84.
Endler JA (1977). *Geographic Variation, Speciation, and Clines*. Princeton University Press: Princeton, NJ.
Evans BJ, Supriatna J, Andayani N, Setiadi MI, Cannatella DC, Melnick DJ (2003). Monkeys and toads define areas of endemism on Sulawesi. *Evol Int J Org Evol* **57**: 1436–1443.
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* **91**: 506–509.
Forattini OP (1965). *Entomologia Médica*, vol. 3, Universidade de Sao Paulo: Sao Paulo.
Glor RE, Vitt LJ, Larson A (2001). A molecular phylogenetic analysis of diversification in Amazonian *Anolis* lizards. *Mol Ecol* **10**: 2661–2668.
Good JM, Sullivan J (2001). Phylogeography of the red-tailed chipmunk (*Tamias ruficaudus*), a northern Rocky Mountain endemic. *Mol Ecol* **10**: 2683–2695.
Griffiths RC, Tavare S (1994). Ancestral inference in population genetics. *Stat Sci* **9**: 307–319.
Griffiths RC, Tavare S (1996). Monte Carlo inference methods in population genetics. *Math Comput Model* **23**: 141–158.
Haffer J (1969). Speciation in Amazonian forest birds. *Science* **165**: 131–137.
Hall JPW, Harvey DJ (2002). The phylogeography of Amazonia revisited: new evidence from Riodinid butterflies. *Evolution* **56**: 1489–1497.
Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**: 95–98.
Hasegawa M, Kishino H, Yano TA (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. *J Mol Evol* **22**: 160–174.
Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.
Hillis DM, Moritz C, Mable BK (eds) (1996). *Molecular Systematics*. Sinauer Associates Inc.: Sunderland, MA.
Hundertmark KJ, Shields GF, Udina IG, Bowyer RT, Danilkin AA, Schwartz CC (2002). Mitochondrial phylogeography of moose (*Alces alces*): late Pleistocene divergence and population expansion. *Mol Phylogenet Evol* **22**: 375–387.
Ibrahim KM, Nichols RA, Hewitt GM (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* **77**: 282–291.
Knowles LL (2001). Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. *Mol Ecol* **10**: 691–701.
Knowles LL, Maddison WP (2002). Statistical phylogeography. *Mol Ecol* **11**: 2623–2635.
Kuhner MK, Yamato J, Felsenstein J (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**: 429–434.
Kumar S, Tamura K, Jakobsen I, Nei M (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244–1245.
Lessa EP, Cook JA, Patton JL (2003). Genetic footprints of demographic expansion in North America, but not Amazonia, during the Late Quaternary. *Proc Natl Acad Sci USA* **100**: 10331–10334.
Mallarino R, Bermingham E, Willmott KR, Whinnett A, Jiggins CD (2005). Molecular systematics of the butterfly genus *Ithomia* (Lepidoptera: Ithomiinae): a composite phylogenetic hypothesis based on seven genes. *Mol Phylogenet Evol* **34**: 625–644.
Masta SE, Laurent NM, Routman EJ (2003). Population genetic structure of the toad *Bufo woodhousii*: an empirical assessment of the effects of haplotype extinction on nested cladistic analysis. *Mol Ecol* **12**: 1541–1554.
Mirabello L, Conn JE (2006). Molecular population genetics of the malaria vector *Anopheles darlingi* in Central and South America. *Heredity* **96**: 311–321.
Moritz C, Faith DP (1998). Comparative phylogeography and the identification of genetically divergent areas for conservation. *Mol Ecol* **7**: 419–429.
Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.
Nei M, Li W (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* **76**: 5269–5273.
Panchal M, Beaumont MA (2007). The automation and evaluation of nested clade phylogeographic analysis. *Evolution* **61**: 1466–1480.
Pedro PM (2004). The impact of habitat fragmentation on a forest-exclusive species of Sabethes mosquito. PhD thesis, University of Leeds: Leeds, UK.
Petit RJ (2008). The coup de grace for the nested clade phylogeographic analysis? *Mol Ecol* **17**: 516–518.
Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
Posada D, Crandall KA (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol* **16**: 37–45.
Posada D, Crandall KA, Templeton AR (2000). GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol Ecol* **9**: 487–488.
Prance GT (1982). *Biological Diversification in the Tropics*. Columbia University Press: New York.
Ramos-Onsins S, Rozas J (2002). Statistical properties of new neutrality tests against population growth. *Mol Bio Evol* **19**: 2092–2100.
Reynolds J, Weir BS, Cockerham CC (1983). Estimation of the co-ancestry coefficient basis for a short-term genetic distance. *Genetics* **105**: 767–779.
Rogers AR, Harpending H (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9**: 552–569.
Rousset F (2000). Genetic differentiation between individuals. *J Evol Biol* **13**: 58–62.
Rozas J, Sanchez-DelBarrio J, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
Sambrook J, Fritsch EF, Maniatis T (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn, Cold Spring Harbor Press: New York.
Schneider S, Roessli D, Excoffier L (2000). *Genetics and Biometry Laboratory*. University of Geneva: Switzerland.
Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994). Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann Entomol Soc Am* **87**: 651–701.
Slatkin M, Hudson RR (1991). Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.
Swofford DL (1993). PAUP-a computer program for phylogenetic inference using maximum parsimony. *J Gen Physiol* **102**: A9.
Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
Tamura K, Nei M (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512–526.

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

Templeton AR (2004). Statistical phylogeography: methods of evaluating and minimizing inference errors. *Mol Ecol* **13**: 789–809.

Templeton AR, Routman E, Phillips CA (1995). Separating population structure from population history—a cladistic analysis of the geographical distribution of mitochondrial-DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**: 767–782.

Thompson JD, Higgins DG, Gibson TJ (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**: 4673–4680.

Vane-Wright RI, Humphries CJ, Williams PH (1991). What to protect—systematics and the agony of choice. *Biol Conserv* **55**: 235–254.

Watterson GA (1975). Number of segregating sites in genetic models without recombination. *Theor Popul Biol* **7**: 256–276.

William J, Ballard O, Kreitman M (1995). Is mitochondrial-DNA a strictly neutral marker. *Trends Ecol Evol* **10**: 485–488.

Zheng XG, Arbogast BS, Kenagy GJ (2003). Historical demography and genetic structure of sister species: deermice (*Peromyscus*) in the North American temperate rain forest. *Mol Ecol* **12**: 711–724.

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