ELSEVIER

Linking phylogenetics with population genetics to reconstruct the geographic origin of a species ☆

Matthew D. Dean* and J. William O. Ballard

University of Iowa, 202 Biology Building, Iowa City, IA 52242, USA

Abstract

Reconstructing ancestral geographic origins is critical for understanding the long-term evolution of a species. Bayesian methods have been proposed to test biogeographic hypotheses while accommodating uncertainty in phylogenetic reconstruction. However, the problem that certain taxa may have a disproportionate influence on conclusions has not been addressed. Here, we infer the geographic origin of *Drosophila simulans* using 2014 bp of the *period* locus from 63 lines collected from 18 countries. We also analyze two previously published datasets, *alcohol dehydrogenase related* and *NADH:ubiquinone reductase 75 kDa subunit precursor*. Phylogenetic inferences of all three loci support Madagascar as the geographic origin of *D. simulans*. Our phylogenetic conclusions are robust to taxon resampling and to the potentially confounding effects of recombination. To test our phylogenetically derived hypothesis we develop a randomization test of the population genetics prediction that sequences from the geographic origin should contain more genetic polymorphism than those from derived populations. We find that the Madagascar population has elevated genetic polymorphism relative to non-Madagascar sequences. These data are corroborated by mitochondrial DNA sequence data.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Drosophila simulans; *period*; Bayesian ancestral reconstruction

1. Introduction

Identifying the geographic origin of a species is an important step towards understanding evolutionary changes over time and space. Traditionally, ancestral inference has proceeded by mapping geographic characters onto a well-resolved phylogenetic reconstruction and inferring the location of the most recent common ancestor and the most parsimonious series of colonization. Increasingly, such a traditional framework has been criticized because full faith must be placed in a single phylogenetic reconstruction, thus failing to incorporate phylogenetic uncertainty. Furthermore, alternative hypotheses are often inadequately evaluated. The goal of this study is to combine phylogenetic and population genetic tools to infer the region of endemicity of the common human commensal fly *Drosophila simulans*.

Huelsenbeck and Imennov (2002) employed Bayesian inference and randomized sampling algorithms to reconstruct the geographic origin of humans. They used data from the pioneering study of Vigilant et al. (1991), which suggested an African origin for 189 human mitochondrial DNA sequences. The Vigilant et al. (1991) study was criticized for biased sampling and inadequate exploration of alternative hypotheses (Maddison et al., 1992; Templeton, 1993). Rather than relying on a single topology, Huelsenbeck and Imennov (2002) explored a set of likely topologies. Such methods may be used to compare alternative hypotheses before and after data are observed (Lewis, 2001). After reanalyzing these data in a Bayesian framework, Huelsenbeck and Imennov (2002) found that support for an ancestral location in Africa increased roughly four times after observing data, which was taken as positive support for the out-of-Africa hypothesis.

---

* Corresponding author. Fax: 1-319-353-3620.
E-mail address: mdean@roosevelt.edu (M.D. Dean).

While these new methods offer many advantages, they are not necessarily immune to the peculiarities of any one data matrix. Intuitively, taxa that occupy the most basal positions in phylogenetic reconstructions will have the largest single influence on parsimonious reconstruction of geographic origin. Therefore, one important question remains for such analyses: How sensitive are conclusions from any one data matrix to the particular taxa sampled? This question is addressed here using two main techniques. First, we incorporated a simple permutation-based resampling strategy into Bayesian analysis to measure robustness of our conclusions to particular taxa included. Second, we developed a randomization test of genetic diversity to test the population genetics prediction that the ancestral location harbors significantly higher genetic diversity than derived populations. We applied these techniques to a worldwide sample of the X-linked *period* locus as well as two previously published datasets from *D. simulans*.

Drosophila simulans is a cosmopolitan species with a broad geographic distribution (http://myweb.uiowa.edu/bballard/map.htm). Based on distributional records and chromosomal evidence, it has been hypothesized that *D. simulans* originated on the island of Madagascar while the closely related *Drosophila melanogaster* originated in continental west Africa (Lachaise et al., 1988). Surprisingly, few datasets have included taxa collected from Madagascar. Rather, a common finding is that continental African populations of *D. simulans* contain more genetic polymorphism than non-African populations (Andolfatto, 2001; Begun and Aquadro, 1993; Hamblin and Veuille, 1999), although the pattern does not hold for all loci (Andolfatto, 2001; Begun et al., 2000). Here we found significantly elevated genetic variation in samples collected from Madagascar, supporting the hypothesis of Lachaise et al. (1988) that *D. simulans* spread out of this African island. This hypotheis is corroborated by mitochondrial DNA data (Ballard, 2004).

2. Materials and methods

Drosophila simulans has been the focus of a variety of molecular evolutionary studies (Akashi, 1996; Andolfatto, 2001; Andolfatto and Kreitman, 2000; Ballard, 2000a; Ballard et al., 2002, 1996; Begun and Whitley, 2000; Coyne et al., 1994; Davis et al., 1996; Hamblin and Veuille, 1999; Hasson et al., 1998; Irvin et al., 1998; Kliman et al., 2000; Rosato et al., 1994; Zurovcova and Eanes, 1999). In this study, we included 63 *D. simulans* lines from 18 countries. To infer the geographic origin of the species, sequences were then grouped into eight geographical locations: continental Africa, Australia, Europe, Madagascar, North America, Oceania, Reunion, and South America (Table 1). At the time of this study lines from Asia were not available.

2.1. Sequence data

We gathered 2014 bp of data from the *period* locus from a worldwide sample of 63 *D. simulans* isofemale lines (GenBank AY575784–AY575846). The same sequence was gathered from the *D. melanogaster* line Oregon R for use as an outgroup (GenBank AY575847). The *period* locus was selected for this study because Kliman and Hey (1993) showed the *period* locus was consistent with the null hypothesis of the MK test (McDonald and Kreitman, 1991), suggesting it provides a phylogenetic signal that is unaffected by strong selection (but see Rosato et al., 1994). Each *period* sequence was gathered from the same individual male to avoid creating mosaic sequences (the *period* locus occurs at 3B1-2 on the X-chromosome of *D. melanogaster*).

DNA was isolated from a single male and DNA amplified following Ballard (2000b). PCRs and sequencing followed Dean et al. (2003). All bases in the dataset were sequenced from both strands.

To test the generality of results inferred from the *period* locus, we included two previously published datasets: *NADH:ubiquinone reductase 75 kDa subunit precursor* (*nd75*) (Ballard et al., 2002) and *alcohol dehydrogenase related* (*adhr*) (Ballard et al., 2002). In *Drosophila*, *nd75* occurs at cytogenetic map position 7D20-22 on the X-chromosome while *adhr* occurs at 35B2 on chromosome 3 of *D. melanogaster*. Both studies sampled 22 *D. simulans* lines: two from Africa, two from North America, seven from Madagascar, six from Oceania, and five from Reunion (the highlighted taxa in Table 1). The *D. melanogaster* Oregon R was included in both datasets. We do not include data from the mitochondrial genome because it is probably strongly influenced by the α-proteobacteria *Wolbachia* (Ballard et al., 1996). Data from mitochondrial DNA suggest that Madagascar or continental east Africa may be the geographic origin of *D. simulans* (Ballard, 2004).

2.2. Bayesian inferences

Genealogical relationships were inferred using Bayesian methods implemented in MrBayes version 3.0B4 (Huelsenbeck and Ronquist, 2001). All sites were included except those containing indels. This program uses a Metropolis-coupled Markov Chain Monte Carlo (MCMCMC) sampling regime to move through possible genealogical trees. MCMCMC begins at a random topology in the sea of possible genealogies. The likelihood of the data given that initial topology is calculated and alterations to the topology are proposed. The likelihood of the data given the newly altered topology is then calculated and compared to the previous likelihood. The new topology is chosen to replace the previous one with a probability dependent on the improvement in likelihood. This chain is iterated many times with the goal of

Table 1
The 63 isofemale lines from which 2014 bp of *period* were sequenced

Locality	Region	mtDNA	*Wolbachia*	Line name	Year
Australia	Australia	*si*II	*w*-	AU117	1999
		*si*II	*w*Au	COFFSA	<1998
Cameroon	Africa	*si*II	*w*-	Y02	1998
		*si*II	*w*Au	Y12	1998
Congo	Africa	*si*II	*w*Ri	CONG02	<1998
		*si*II	*w*Ri	CONG11	<1998
Ecuador	S. America	*si*II	*w*-	EC101	2000
		*si*II	*w*-	EC128	2000
		*si*II	*w*Au	EC125	2000
		*si*II	*w*Au	EC126	2000
		*si*II	*w*Ri	EC131	2000
		*si*II	*w*Ri	EC132	2000
France	Europe	*si*II	*w*Ri	53Val	1993
		*si*II	*w*Ri	54Vill	1992
Greece	Europe	*si*II	*w*-	GR134	2000
		*si*II	*w*Ri	GR100	2000
Hawaii	Oceania	*si*I	*w*Ha	**HW00**	1990?
		*si*I	*w*Ha	**HW09**	1998
Jamaica	S. America	*si*II	*w*Au	JM001	2000
		*si*II	*w*Au	JM002	2000
Japan	Oceania	*si*II	*w*-	SCJ9319	1993
		*si*II	*w*Ri	SCJ9316	1993
Kenya	Africa	*si*II	*w*-	SL61	1979
		*si*II	*w*Ri	**C167**	1973
Madagascar	Madagascar	*si*II	*w*-	MD238	1998
		*si*II	*w*Au	**MD106**	1998
		*si*II	*w*Au	**MD225**	1998
		*si*III	*w*-	**MD111**	1998
		*si*III	*w*-	**MD221**	1998
		*si*III	*w*-	**MDW86**	<1993
		*si*III	*w*Ma	**MD112**	1998
		*si*III	*w*Ma	**MD199**	1998
New Caledonia	Oceania	*si*I	*w*-	NC103	1999
		*si*I	*w*Ha	NC112	1999
		*si*I	*w*Ha	NC115	1999
		*si*I	*w*Ha+*w*Ma	**NC037**	1991
		*si*I	*w*Ha+*w*Ma	**NC048**	1991
		*si*I	*w*Ha+*w*Ma	NC102	1999
		*si*I	*w*Ma	NC117	1999
		*si*I	*w*Ma	NC141	1999
Reunion	Reunion	*si*II	*w*-	**RU000**	1993
		*si*III	*w*-	**RU001**	1998
		*si*III	*w*-	**RU035**	1998
		*si*III	*w*-	**RU259**	1979
		*si*III	*w*Ma	RU005	1998
		*si*III	*w*Ma	**RU007**	1998
		*si*III	*w*Ma	RU008	1998
Seychelles	Oceania	*si*I	*w* Ha+*w*Ma	SC01	<1998
		*si*I	*w*Ha+*w*Ma	SC02	<1998
		*si*II	*w*Ri	SC00	<1998
South Africa	Africa	*si*II	*w*-	SA5	<1998
		*si*II	*w*Ri	SA1	2000
Tahiti	Oceania	*si*I	*w*Ha	**TT00**	1998
		*si*I	*w*Ha	**TT01**	1998

Table 1 (*continued*)

Locality	Region	mtDNA	*Wolbachia*	Line name	Year
Tanzania	Africa	*si*III	*w*Ma	KC9A	1997
Tunisia	Africa	*si*II	*w*-	TUN50	<1998
		*si*II	*w*Ri	TUN37	<1998
USA	N. America	*si*II	*w*-	DSW	1985
		*si*II	*w*-	LA44	1994
		*si*II	*w*Au	LA02	1994
		*si*II	*w*Ri	DSR	1987
Zimbabwe	Africa	*si*I	*w*-	Z03	1994
		*si*II	*w*Ri	Z26	1994

Highlighted taxa are those for which *adhr* and *nd75* data were available from Ballard (2000a) and Ballard et al. (2002), respectively.

exploring a relatively likely set of topologies. To avoid the analysis becoming trapped in local optima, four chains were run, three of which were "heated" with $\beta = 0.2$, where heat $= 1/(1 + \beta * (ID - 1))$ and ID is 1, 2, 3, or 4.

Implementing an optimal model of sequence evolution is important in any phylogenetic analysis. Using likelihood ratio tests (Huelsenbeck and Crandall, 1997; Swofford et al., 1996), we determined that a general time reversible model with gamma distributed rate variation among sites and a proportion of invariant sites (GTR+Γ+I) was best for the *period* data. All sites except indels were included. Using the GTR+Γ+I model, 1,000,000 generations of MCMCMC were run, sampling every 100th topology and calculating its likelihood. This MCMCMC sampling regime produced 10,000 trees sampled in accordance to their likelihoods. The first 500 trees were discarded as the burn-in because they were sampled early and thus relatively unlikely. Calculating the number of times a particular clade appeared in the remaining 9500 trees assessed confidence of clades. Maximum likelihood estimates of the parameters of the GTR+Γ+I model, as well as their 95% credible intervals, were calculated from the posterior distribution of topologies.

All conclusions were robust to the number of MCMCMC generations run; if we analyze the last 95,000 trees in the posterior distribution of a 10,000,000-generation (instead of 1,000,000) MCMCMC, no conclusions change.

2.3. Ancestral reconstructions

A priori, it may be expected that geographic regions represented by more sequences will have a higher probability of being reconstructed as the ancestral location by chance. To calculate such prior probabilities, 9500 random topologies were generated assuming a uniform prior, then sequences were coded by geographic region, and ancestral regions reconstructed using accelerated transformation parsimony. The frequency that a particular region was mapped to the most recent common ancestor of *D. simulans* was calculated for both prior and posterior distributions of trees. An alternative procedure might utilize coalescence theory (Beerli and Felsenstein, 1999).

By comparing prior to posterior probabilities, the change in opinion of each geographic hypothesis *before* and *after* observing data was quantified. This change in opinion is referred to as the Bayes factor and can be directly compared among competing hypotheses given a particular data matrix. Huelsenbeck and Imennov (2002) suggested that in the case of multiple datasets, one could use the posterior probability of one analysis as the prior probability for subsequent analyses. We do not take this strategy because *nd75* and *adhr* do not include samples from Australia, Europe, and South America and only two samples are from continental Africa (Table 1).

2.4. Recombination

An assumption of Bayesian analyses is that the locus of interest has not undergone recombination. With recombination, different parts of the locus may have different histories, violating the assumption that phylogenetic history can be represented with a bifurcating tree. It should be noted, however, that lack of recombination violates the assumption that all sites are evolving independently. For the purpose of this paper, we are interested in whether or not recombination compromises our conclusions of ancestral reconstruction.

For all three datasets, we tested for recombination within *D. simulans* using the four gamete test (Hudson and Kaplan, 1985) as implemented in DNAsp version 3.99 (Rozas and Rozas, 1997). This test relies on the assumption that the only way to observe all four possible phase combinations from two segregating sites is if a recombination event occurred between those two sites. The four gamete test is therefore taken as a minimum number of recombination events, but it should be noted that multiple substitutions could produce the four possible gametes without recombination.

We tested whether recombination altered our phylogenetic conclusions using PLATO (Grassly and Holmes, 1997). PLATO uses a sliding window approach to determine if the average per-site likelihood of a given window of DNA is significantly low compared to the average per-site likelihood of the overall dataset. One explanation for significantly conflicting regions is that recombination has concatenated windows of sequence that have conflicting evolutionary histories. An alternative explanation is that the rate of evolution is heterogeneous along a sequence (Grassly and Holmes, 1997; Schierup and Hein, 2000). A general time reversible model with Γ-distributed rate variation among sites was employed in PLATO. For each locus, the likelihood estimates of the rate matrix, base frequencies, and Γ shape parameter were estimated using PAUP* version 4.0b10 (Swofford, 1993), then implemented in PLATO. Preliminary runs indicated that a minimum window size of 100 yielded the most informative results. PLATO does not accept polytomies in the input tree, so trees were arbitrarily resolved with branch lengths of zero. Calculated likelihoods are equivalent whether or not polytomies are arbitrarily resolved with branch lengths of zero.

2.5. Randomization test of genetic diversity

Phylogenetic investigations supported an out-of-Madagascar hypothesis (see below). This hypothesis can be corroborated or rejected by population genetic theory, which predicts that Malagasy populations will have increased polymorphism compared to non-Malagasy populations (Wares and Cunningham, 2001). This prediction assumes that admixture from divergent populations has not occurred in Madagascar, and selection has not acted to remove variation from ancestral populations and/or promote variation in derived populations. To test this prediction, π (Nei and Li, 1979) was estimated from the entire dataset and the Madagascar sequences only. Then, an equal number of non-Madagascar sequences were randomly chosen from the entire dataset, and π was recalculated. This step was iterated 10,000 times to ask if π from Malagasy flies was significantly greater than that from non-Malagasy flies. This randomization test was also performed for regions identified by PLATO as anomalous. As we have a specific prediction this is a one-tailed test.

2.6. Tests of neutrality

Violation of neutrality may indicate that selection is operating on the loci under question or that population size has changed in the detectable past. Under a neutral model of evolution, π and θ (Watterson, 1975), calculated from the total number of mutations, should be approximately equal. Tajima's D (1989) tests this neutral

prediction. Following a β distribution, the neutral expectation is that D should be close to zero (Tajima, 1989). A significantly negative D indicates selection prevented mutations from increasing in frequency and/or that the population has been expanding. So that the recombination parameter R (Hudson, 1987) could be included, statistical significance of D was determined using coalescent simulations implemented in DNAsp 3.99 (Rozas and Rozas, 1997).

As a further test of the neutral equilibrium model of molecular evolution we employed the MK test (McDonald and Kreitman, 1991). This tested the neutral prediction that the ratio of polymorphic and fixed nonsynonymous substitutions was proportional to that from synonymous sites.

3. Results

3.1. Sequence data

In the *period* dataset, there were 218 segregating sites. Within *D. simulans*, 10 of 165 segregating sites were nonsynonymous mutations. There were 32 distinct haplotypes and the GC content was 64% (63% in coding regions and 55% in noncoding regions). To investigate the variation within *period* we tested for an association between the number of segregating sites and the length of the three introns and four exons (Fig. 1). Contingency table analyses showed that the proportion of segregating sites differed significantly among regions ($\chi_6^2 = 58.99$, $P < 0.001$). As expected, this result was due to the higher proportion of segregating sites in noncoding regions (46 segregating sites from 196 bp total) compared
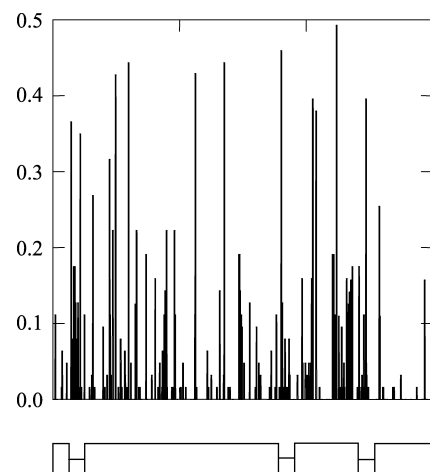

Fig. 1. The 165 segregating sites found within *D. simulans*, plotted along a schematic representation of exons (open boxes) and introns (lines) of the 2014 bp region sequenced from the *period* locus. Shown on the *y*-axis is the frequency of the least frequent mutation per segregating site.

to coding regions (119 segregating sites from 1818 bp). There was no difference among the three introns ($\chi_2^2 = 0.08$, $P = 1$), but a deficiency of segregating sites in the fourth exon ($\chi_3^2 = 10.41$, $P = 0.03$ for all four exons and $\chi_2^2 = 1.02$, $P = 1.0$ when the fourth exon was removed). Within *D. simulans* there were three indels. One occurred in intron 1 (22 taxa had a T insertion at site No. 146) and two occurred in intron 2 (in one case five taxa had a C insertion at position No. 1252 and in the other one a taxon from Madagascar had a T deletion at site No. 1215). A table of polymorphic sites for the *period* locus is presented as supplementary material (MPE website).

In the *nd75* dataset, there were 165 segregating sites. Within *D. simulans*, 3 of 121 segregating sites were nonsynonymous mutations. There were 16 distinct haplotypes and the GC content was 59.3% (62.2% in coding

regions and 46.0% in noncoding regions). We tested for an association between the number of segregating sites and the length of the four introns and four exons. Contingency analysis showed that the proportion of segregating sites differed significantly among regions ($\chi_7^2 = 67.4$, $P < 0.001$). As expected, this result was due to the higher proportion of segregating sites in noncoding regions (46 segregating sites from 420 bp total) compared to coding regions (75 segregating sites from 1938 bp total). There was no difference among the four exons ($\chi_3^2 = 6.97$, $P = 0.07$), but an excess of segregating sites in the fourth intron ($\chi_3^2 = 16.3$, $P < 0.001$ for all four introns and $\chi_2^2 = 2.32$, $P = 0.31$ when the fourth intron was removed). Within *D. simulans* there was one indel (two taxa had a TC insertion at site Nos. 1948–1949).

The *adhr* dataset consisted of a single intron containing 39 segregating sites. There were 18 distinct
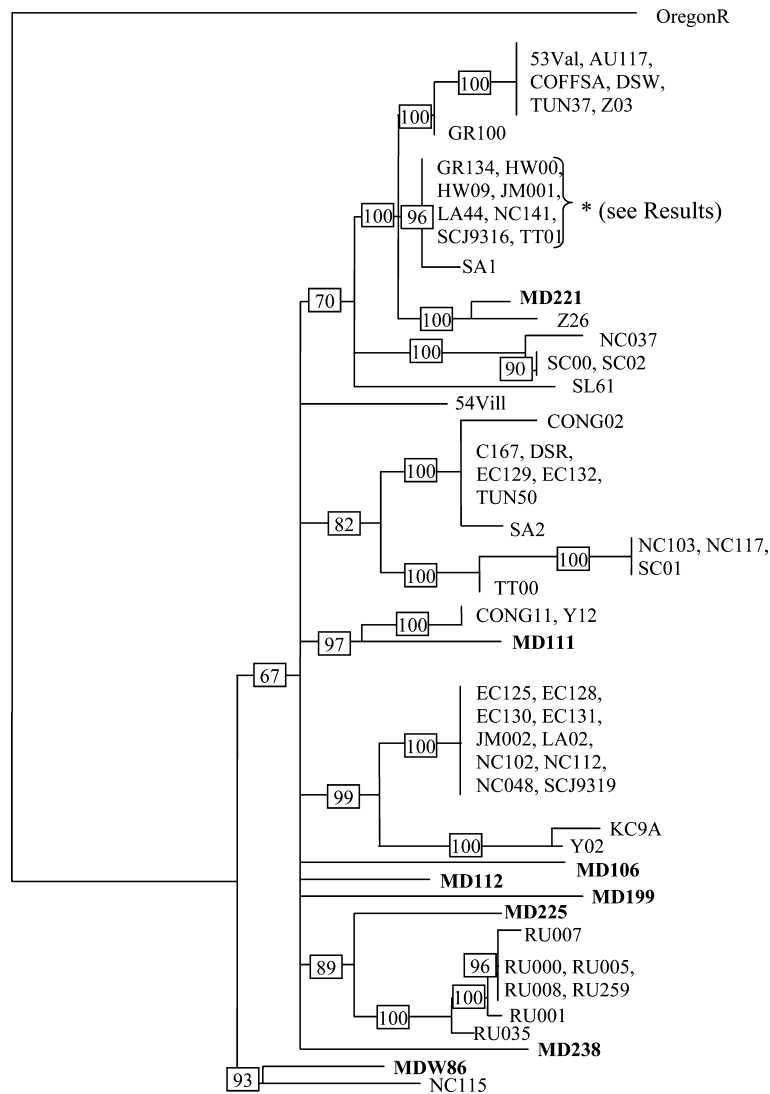

Fig. 2. Genealogical inferences based on the *period* locus. A majority rule consensus tree was drawn from 9500 trees (10,000–500 burnin) in the posterior distribution. Sequences separated by commas were identical. Boxed numbers on branches indicate posterior probability support for clades. Bold indicates sequences from Madagascar. These genealogical relationships remained unchanged if we ran 10,000,000 generations of MCMCMC instead of 1,000,000 generations. Note: *D. melanogaster* is the outgroup.

haplotypes and the GC content was 38.0%. Within *D. simulans* there were nine indels.

3.2. Bayesian inferences

The genealogical relationships inferred using Bayesian analysis of the *period* locus are shown in Fig. 2, and the parameters estimated from the $G+\Gamma+I$ model in Table 2. The $C \leftrightarrow T$ transition is the dominant molecular change.

Table 2
Parameters of the $G+\Gamma+I$ model estimated from Bayesian inference of *period*

Parameter	Mean	95% credible interval	
r(G T)	1.00	1.00	1.00
r(C T)	14.17	6.83	29.47
r(C G)	1.59	0.71	3.20
r(A T)	8.39	3.66	15.45
r(A G)	9.52	4.14	18.59
r(A C)	1.35	0.55	2.97
Freq(A)	0.21	0.19	0.23
Freq(C)	0.32	0.30	0.34
Freq(G)	0.31	0.29	0.33
Freq(T)	0.16	0.15	0.18
α	0.05	0.05	0.05
P(invariant)	0.70	0.68	0.73

There was a strong correlation between posterior probabilities of clades in the two runs ($F = 128.34$, $P < 0.001$), and the chains appeared to have converged upon a region of likelihood scores after about 100 generations. Conclusions do not change if we run the chains for 10,000,000 generations.

The sequences collected from Reunion all formed a strongly supported monophyletic group (Fig. 2). This clade was defined by a unique nonsynonymous mutation. Aside from this pattern, there was little geographic structure (Fig. 2). For example, the same sequence was found in Greece, Hawaii, Jamaica, Florida, New Caledonia, Japan, and Tahiti (clade with asterisk, Fig. 2). This result suggested long range dispersal of *period* alleles is not unusual in this species and that migration may have influenced our conclusions. We evaluate this possibility in the discussion.

Three inferences from the *period* data suggest that Madagascar was the geographic origin of *D. simulans*. First, all sequences from flies collected in Madagascar were unique (sample names preceded by MD and in bold, Fig. 2). Similarly, Derome et al. (2004) sampled 697 bp of the *Vermilion* locus from 15 Madagascar flies, and found that all 15 sequences were unique. Second, branches leading to Madagascar sequences were quali-
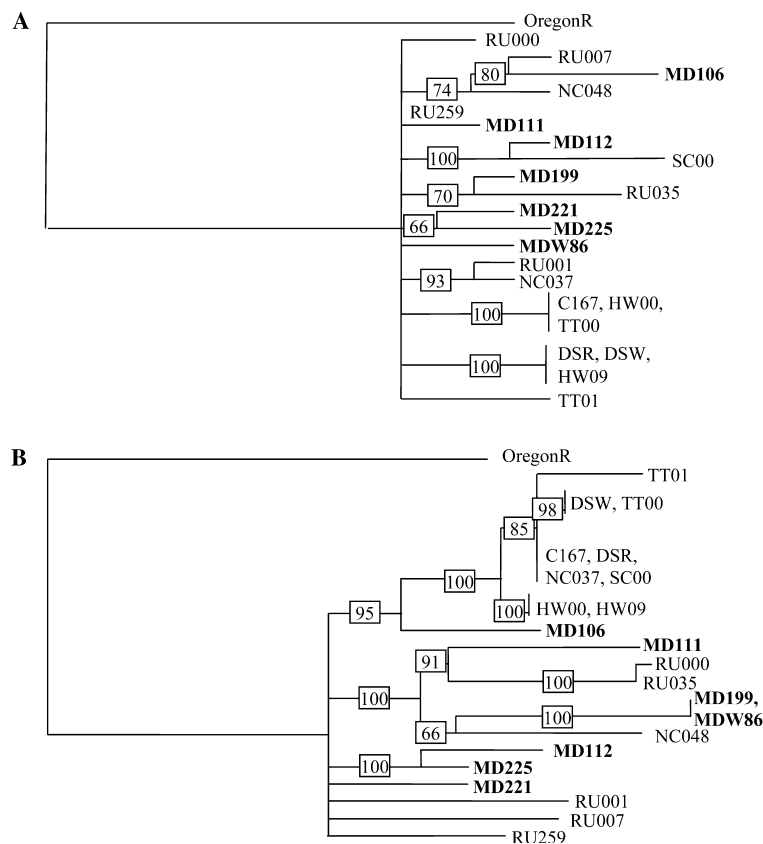

Fig. 3. Genealogical inferences based on the (A) *nd75* and (B) *adhr* loci. In each case, a majority rule consensus tree was drawn from 9500 trees (10,000–500 burnin) in the posterior distribution. Sequences separated by commas were identical. Boxed numbers on branches indicate posterior probability support for clades. Bold indicates sequences from Madagascar. Note: *D. melanogaster* is the outgroup.

tatively longer than those leading to non-Madagascar sequences. Longer branches may occur from increased autapomorphies, from increased homoplasies, or from a combination of the two. Third, Madagascar sequences seemed to occupy basal positions more frequently than non-Madagascar sequences.

Qualitatively, the two additional datasets supported the out-of-Madagascar hypothesis for *D. simulans*. The *nd75* and *adhr* phylogenies also showed a general lack of geographic structure but Madagascar sequences again occurred in basal positions (Fig. 3). In the following section, we quantified the probability that Madagascar is the geographic origin of *D. simulans* and compared this probability to competing hypotheses.

3.3. Ancestral reconstructions

For the *period* dataset we reconstructed the ancestral region of 9500 trees in the posterior distribution (10,000 trees generated during MCMCMC sampling minus 500 "burnin" trees) as well as the prior distribution. A Bayes factor of 24.30 quantified the change in opinion that Madagascar is the ancestral location after observing *period* data (Table 3, row 1), providing "strong" support for a hypothesis (Raftery, 1995). The only other locations with posterior probabilities greater than zero were Europe and Oceania, probably because one sequence from each of these locations (54Vill from Europe and NC115 from Oceania) occupied basal positions in the

genealogy (Fig. 2). In contrast, support for five of the seven alternative hypotheses—continental Africa, Australia, North America, Reunion, and South America—decreased after observing data (Table 3, row 1).

To investigate sensitivity to taxon sampling systematically, we performed a permutation-based resampling study. First, all possible datasets including 62 (63 − 1) *D. simulans* lines were constructed. Next, 100 datasets each were created after randomly removing 3, 5, and 7 *D. simulans* lines (300 datasets total). Because there were only eight sequences from Madagascar, removing eight or more individuals might create datasets with 0 Madagascar sequences, leaving the Madagascar hypothesis undefined. For each dataset, the Bayesian analyses were repeated. Reassuringly, the Bayes factors remained stable through this resampling (Table 3, rows 2–5) and the Madagascar hypothesis was supported by an order of magnitude more than the next most supported hypothesis. Specifically, the Madagascar hypothesis ranged from 23.6 to 24.9, the European hypothesis ranged from 2.2 to 2.5, and the continental Africa hypothesis ranged from 0.2 to 0.5. No other hypotheses received a Bayes factor of greater than zero.

The most extreme test of taxon sensitivity involved the removal of the most basal Madagascar sequence. After removal of MDW86, the Bayes factor decreased from 24.30 to 5.00 (Table 3, row 6). Even in this case Madagascar was still the most favored hypothesis, with

Table 3
Bayes factor support for alternative hypotheses of the ancestral location

Locus/sites	Geographic hypothesis							
	Africa (continental)	Australia	Europe	Madagascar	N. America	Oceania	Reunion	S. America
period								
All[a]	0.20	0.00	1.90	24.30	0.00	0.00	0.00	0.00
1 taxon[b]	0.2 ± 0.0	0.0 ± 0.0	2.3 ± 0.2	24.9 ± 1.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
3 taxa[b]	0.3 ± 0.0	0.0 ± 0.0	2.2 ± 0.2	24.9 ± 2.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
5 taxa[b]	0.3 ± 0.0	0.0 ± 0.0	2.2 ± 0.2	24.7 ± 1.3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
7 taxa[b]	0.5 ± 0.2	0.0 ± 0.0	2.5 ± 0.2	23.6 ± 1.6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
MDW86[c]	1.80	0.00	1.70	5.00	0.00	0.00	0.00	0.00
99–456[d]	1.98	0.00	0.17	6.00	0.05	0.01	0.04	0.45
1514–1696[d]	0.02	0.00	377.74	0.13	0.00	0.00	0.00	0.00
Remaining[e]	1.49	0.01	0.05	7.87	0.00	0.08	0.00	0.00
nd75								
All[a]	0.00			31.72	0.00	0.00	0.17	
1880–1982[d]	0.05	NA[f]		5.67	0.01	0.05	1.27	NA[f]
Remaining[e]	0.12			23.01	0.00	0.01	0.10	
adhr								
All[a]	0.00			1.87	0.28	0.22	1.00	

[a] All taxa/sites included.
[b] Randomly subsampled datasets after removing number of taxa indicated. Numbers are the average ± standard error Bayes factor.
[c] Dataset after excluding taxon MDW86.
[d] Datasets with all taxa included and including only sites indicated, which are regions identified by PLATO as phylogenetically anomalous.
[e] Datasets with all taxa included and all sites included except those indicated in preceding row(s).
[f] These cells are not applicable since the geographic regions indicated were not sampled for those datasets.

its Bayes factor greater than that from continental Africa (= 1.80) and Europe (= 1.70).

The two additional datasets offered independent support for the out-of-Madagascar hypothesis. For both genes, an origin in Madagascar received more support than all other hypotheses. The *nd75* and *adhr* datasets supported an origin in Madagascar with a Bayes factor of 31.72 and 1.87, respectively (Table 3, rows 10 and 13).

3.4. Recombination

In the *period* locus, the four gamete test identified a minimum number of 22 recombination events. PLATO identified two anomalous regions, spanning sites 99–456 and 1514–1696, that were inconsistent with the genealogy derived from the remainder of the locus. Both regions span introns and exons at approximately 50%. Region 99–456 supported the Madagascar hypothesis with a Bayes factor of 6.00. The only other hypothesis supported with a Bayes factor greater than 1 was continental Africa, with a Bayes factor of 1.98 (Table 3, row 7).

Conservatively, the region spanning 1514–1696 supports an ancestral origin in Europe with a Bayes factor of 377.74 (Table 3, row 8). No other hypothesis was supported with a Bayes factor of greater than 1. The high support for a European hypothesis occurred because taxon 54Vill occupied the most basal position of the tree, with all other *D. simulans* uniting as a monophyletic sister group. Two synonymous sites are responsible for this placement. At site No. 1577, 54Vill has an adenine, a state shared with the outgroup Oregon R. All other *D. simulans* taxa have a guanine. At site No. 1562, 54Vill is the only taxon to have a cytosine, while all other *D. simulans* have either a guanine or a thymine. Removing sites 1562 and 1577 decreased support for a European hypothesis from a Bayes factor of 377.74 to 1.56 and increased support for a Madagascar hypothesis to 7.32. After removing 54Vill from the analysis entirely, PLATO still identified region 1514–1696 as anomalous, but supported the Madagascar hypothesis with a Bayes factor of 43.24.

After removing the two anomalous regions identified by PLATO, the remaining *period* data supported the Madagascar hypothesis with a Bayes factor of 7.87 (Table 3, row 9). The only other hypothesis that was supported with a Bayes factor greater than 1 was continental Africa, which had a Bayes factor of 1.49.

In *nd75*, the four gamete test identified a minimum of 20 recombination events. PLATO identified one anomalous region, spanning from position 1880 to 1982. This region supported the Madagascar hypothesis with a Bayes factor of 5.67 and no other hypothesis was supported with a Bayes factor greater than 1 (Table 3, row 11). The data remaining after removal of this anomalous region supported the Madagascar hypothesis with a Bayes factor of 23.01 (Table 3, row 12).

In *adhr*, the four gamete test identified a minimum of five recombination events. PLATO did not identify any anomalous regions using minimum window sizes of 50 or 100.

3.5. Randomization test of genetic diversity

If Madagascar were the ancestral location, theory predicts that flies collected here should possess greater nucleotide diversity than flies collected from other regions. This result was observed for all three loci (Table 4). For the eight *period* sequences from Madagascar, $\pi = 0.014$. This was greater than 9940 of the 10,000 iterations ($P < 0.01$) where eight non-Madagascar sequences were randomly chosen (average $\pi = 0.011 \pm 0.001$ standard deviation). For the anomalous region spanning 99–456, $\pi = 0.024$ for Madagascar sequences, which was greater than 9725 of the 10,000 iterations. For anomalous region 1514–1696, $\pi = 0.028$ for the Madagascar sequences, which was greater than 6138 of the 10,000 iterations. For the region remaining after removal of the two anomalous regions, $\pi = 0.011$ from Madagascar sequences, which was greater than 9943 of the 10,000 iterations. The smallest regions have the least power to reject the null hypothesis, nevertheless all results are in the direction predicted if Madagascar were the ancestral geographic origin.

The *nd75* dataset consisted of 22 *D. simulans* lines, seven of which were collected from Madagascar (Table 1). The *nd75* sequences from Madagascar had $\pi = 0.014$. This was greater than 9991 of the 10,000 iterations ($P < 0.01$) where seven non-Madagascar sequences were randomly chosen ($\pi = 0.009 \pm 0.002$). For the anomalous region spanning 1880–1982, $\pi = 0.045$ from Madagascar sequences, which was greater than 5629 of the

Table 4

Probability that Madagascar sequences hold more polymorphism than an equal number of non-Madagascar sequences, determined with 10,000 iterations of a randomization test

Locus/ sites	P(Madagascar > non-Madagascar)
period	
All[a]	0.99
99–456[b]	0.97
1514–1696[b]	0.61
Remaining[c]	0.99
nd75	
All[a]	1.00
1880–1982[b]	0.56
Remaining[c]	1.00
adhr	
All[a]	0.69

[a] All taxa/sites included.

[b] Datasets with all taxa included and including only sites indicated, which are regions identified by PLATO as phylogenetically anomalous.

[c] Datasets with all taxa included and all sites included except those indicated in preceding row(s).

10,000 iterations. For the region remaining after removal of this anomalous region, $\pi = 0.012$ from Madagascar sequences, which was greater than all 10,000 randomized iterations. In this case, only the smallest region cannot reject the null hypothesis.

The *adhr* dataset contained the same taxa as the *nd75* dataset. The sequences from Madagascar had $\pi = 0.015$ and seven randomly chosen non-Madagascar sequences had $\pi = 0.014 \pm 0.002$. The deviation of Madagascar π compared to non-Madagascar π was not significant but in the direction predicted by theory (6946 of the 10,000 iterations, $P = 0.30$).

3.6. Tests of neutrality

In general, we observed a significantly negative Tajima's D but not a rejection of the MK test. A significantly negative Tajima's D suggests the population has expanded and/or purifying selection prevented mutations from increasing in frequency. The MK test identifies selection by comparing the ratio of polymorphic and fixed nonsynonymous substitutions to the same ratio from synonymous sites. Taken together, our observations are logically consistent with the idea that *D. simulans* is a human commensal and the population has been expanding.

Using coalescent simulations with the empirically estimated recombination parameter ($R = 63.3$), Tajima's D from the entire *period* dataset was significantly negative ($D = -1.24$, $P < 0.01$). For the anomalous region spanning sites 1514–1696, Tajima's D is not significantly negative. This is probably due to the loss of power in such a small region of DNA, because both the anomalous region 99–456 and the region remaining after removal of the two anomalous regions maintained their statistical significance.

For *nd75*, Tajima's D was significantly negative ($D = -0.76$, $P < 0.05$), as determined with coalescent simulations and the empirically determined recombination rate ($R = 31.6$). The anomalous region spanning 1880–1982 did not display a significantly negative Tajima's D, nor did the region remaining after removal of this anomalous region. For *adhr*, Tajima's D was significantly negative ($D = -1.54$, $P < 0.01$), based on coalescent simulations and the empirically determined recombination rate ($R = 2.3$). Tajima's D test results from all three datasets did not change after analyzing Madagascar and non-Madagascar sequences separately.

The null hypothesis of the MK test was not violated for either *period* or *nd75* ($P > 0.50$). The results of the MK test did not change after analyzing anomalous regions separately. Results also did not change after analyzing Madagascar and non-Madagascar sequences separately. The *adhr* dataset did not include exons and the MK test was not be performed.

4. Discussion

The distribution of a species may be described in geographic and ecological terms. Geographic distributions are often limited in part by history or by ecological factors such as climate or competition with other species. Reconciling the region of endemicity of a species facilitates understanding the processes that have shaped its long-term evolution. It will also help us understand the evolution of specific morphological and molecular characters. *D. simulans* is a human commensal that has colonized North America and Europe recently (Lachaise et al., 1988). Bayesian inference strongly supported the hypothesis that *D. simulans* originated in Madagascar, a conclusion that was robust to taxon sampling, recombination, or the datasets analyzed. One anomalous region gave high support for a European hypothesis, but this appears to be due to the state distribution of two synonymous sites. Consistent with the subsequent population genetics prediction, sequences from Madagascar harbored more genetic polymorphism than non-Madagascar sequences.

Mitochondrial DNA data support the hypothesis that Madagascar is the ancestral site of *D. simulans* (Ballard, 2004). However, the mtDNA genome has clearly been influenced by strong selection (Ballard et al., 1996). *D. simulans* harbors three genetically distinct mitochondrial haplogroups (*si*I, *si*II, and *si*III). These haplogroups differ by ~2% at the nucleotide level, with a significant deficiency of variation found within each group (Ballard, 2000a). The *si*I haplogroup has been found on the Seychelles, New Caledonia, Hawaii, and Tahiti; the *si*III haplogroup on Reunion Island, Madagascar, and continental east Africa; the *si*II haplogroup has been collected nearly everywhere except Hawaii, Tahiti, and New Caledonia (Table 1), and the polar regions. Madagascar is in close proximity to all three haplogroups and is one of four places where two distinct haplogroups exist in sympatry (the others are Reunion Island, Seychelles and continental east Africa).

We will now consider alternate explanations for the data. Of the three loci analyzed, *period* and *nd75* give strongest support for an out-of-Madagascar hypothesis. The Bayes factors calculated from *period* and *nd75* were an order of magnitude higher than *adhr*. Malagasy flies had significantly higher genetic polymorphism than non-Malagasy flies for *period* and *nd75* but not for *adhr*. *period* and *nd75* both occur on the X-chromosome, while *adhr* occurs on the third chromosome. Therefore, one alternative explanation is that support for the out-of-Madagascar hypothesis was specific to the X-chromosome. For this to be true, we must posit that demographic or selective histories differed among X-linked and autosomal genes possibly occurring by differential migration or survival of males and females. The X-chromosome is expected to have a smaller effective

population size than autosomal genes because there are three-fourths as many X-chromosomes as any one autosome, assuming an equal number of effective males and females. To more systematically test if support for the out-of-Madagascar hypothesis is chromosome-specific, sequences from Madagascar should be included in future sampling or added to existing datasets of autosomes.

A second potential bias may have arisen by the inclusion of specific taxa. After removing MDW86 from the *period* dataset Bayesian inference and ancestral reconstruction changed dramatically. MDW86 differs from all other Madagascar sequences in that it was collected prior to 1993 by R. Russell; the other seven Madagascar sequences were all collected in 1998 by this laboratory (Table 1). It is possible that MDW86 has accumulated deleterious mutations in the laboratory, but there are two counterarguments against this possibility. First, MDW86 was not the most basal Madagascar sequence in the *nd75* and *adhr* phylogenies (Fig. 3). Second, several other fly lines, including the two from Kenya (C167 and SL61), were collected prior to 1993 yet did not preferentially occupy basal positions in the phylogeny.

A third potential bias in inferring Madagascar as the ancestral geographic origin may arise if flies immigrated from differentiated populations into Madagascar. Admixture may lead to elevated polymorphism that could be considered ancestral using accelerated transformation parsimony. Under such a scenario, the sequences from Madagascar may actually have originated elsewhere. All Madagascar sequences are unique, suggesting this is not likely given the datasets analyzed. However, Ballard (2004) showed that *Wolbachia w*Au-infected *si*II flies moved from Australia to Madagascar. While admixture may occur in the mitochondrial genome, we saw no evidence of it in the nuclear genome because no Australia sequences cluster with Madagascar sequences (Fig. 2). However, as only two fly lines collected from Australia were included in this study, admixture remains a possibility. Nevertheless, the existing data suggest that admixture does not adequately explain the elevated polymorphism found in Madagascar.

Finally, selection may have removed polymorphism from an ancestral population and/or promoted variation in derived populations. Either scenario offers an alternative hypothesis to the significantly greater polymorphism found in Madagascar. However, selection is likely to act on specific loci or linkage groups and we found elevated genetic diversity in multiple loci. Furthermore, no locus showed evidence of selection based on the MK test. Given the preponderance of evidence, the simplest explanation is that Madagascar has retained ancestral genetic polymorphism and is the geographic origin from which *D. simulans* expanded.

This study documents the need to sample broadly both geographically and genetically. We have found a previously under appreciated source of genetic variation in *D. simulans* from Madagascar. The study also underscores the need to assess the sensitivity of conclusions to specific taxa included in the sample. We address this issue by including multiple datasets, incorporating a resampling scheme into Bayesian analyses, and performing a randomization test of genetic diversity. We propose that critical examination of the evolutionary and ecological forces that have shaped the long-term history of *D. simulans* should include samples from Madagascar.

Acknowledgments

References

Akashi, H., 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144, 1297–1307.

Andolfatto, P., 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18, 279–290.

Andolfatto, P., Kreitman, M., 2000. Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans*. Genetics 154, 1681–1691.

Ballard, J.W.O., 2000a. Comparative genomics of mitochondrial DNA in *Drosophila simulans*. J. Mol. Evol. 51, 64–75.

Ballard, J.W.O., 2000b. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. J. Mol. Evol. 51, 48–63.

Ballard, J.W.O., 2004. Sequential evolution of a symbiont inferred from the host: *Wolbachia* and *Drosophila simulans*. Mol. Biol. Evol..

Ballard, J.W.O., Chernoff, B., James, A.C., 2002. Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans*. Evolution 56, 527–545.

Ballard, J.W.O., Hatzidakis, J., Karr, T.L., Kreitman, M., 1996. Reduced variation in *Drosophila simulans* mitochondrial DNA. Genetics 144, 1519–1528.

Beerli, P., Felsenstein, J., 1999. Maximum-likelihood estimation of migration and effective population numbers in two population using a coalescent approach. Genetics 152, 763–773.

Begun, D.J., Aquadro, C.F., 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365, 548–550.

Begun, D.J., Whitley, P., 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA 97, 5960–5965.

Begun, D.J., Whitley, P., Todd, B.L., Waldrip-Dail, H.M., Clark, A.G., 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics 156, 1879–1888.

Coyne, J.A., Crittenden, A.P., Mah, K., 1994. Genetics of a pheromonal difference contributing to reproductive isolation in *Drosophila*. Science 265, 1461–1464.

Davis, A.W., Roote, J., Morley, T., Sawamura, K., Herrmann, S., Ashburner, M., 1996. Rescue of hybrid sterility in crosses between *D. melanogaster* and *D. simulans*. Nature 380, 157–159.

Dean, M.D., Ballard, K.J., Glass, A., Ballard, J.W.O, 2003. Influence of two Wolbachia strains on population structure of east African *Drosophila simulans*. Genetics 165, 1959–1969.

Derome, N., Métayer, K., Montchamp-Moreau, C., Veuille, M., 2004. Signature of selective sweep associated with the evolution of *sex-ratio* drive in *Drosophila simulans*. Genetics 166, 1357–1366.

Grassly, N.C., Holmes, E.C., 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. 14, 239–247.

Hamblin, M.T., Veuille, M., 1999. Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics 153, 305–317.

Hasson, E., Wang, I.N., Zeng, L.W., Kreitman, M., Eanes, W.F., 1998. Nucleotide variation in the *triosephosphate isomerase* (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 15, 756–769.

Hudson, R.R., 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50, 245–250.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28, 437–466.

Huelsenbeck, J.P., Imennov, N.S., 2002. Geographic origin of human mitochondrial DNA: accommodating phylogenetic uncertainty and model comparison. Syst. Biol. 51, 155–165.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Irvin, S.D., Wetterstrand, K.A., Hutter, C.M., Aquadro, C.F., 1998. Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in New World populations. Genetics 150, 777–790.

Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J., McCarter, J., Wakeley, J., Hey, J., 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics 156, 1913–1931.

Kliman, R.M., Hey, J., 1993. DNA sequence variation at the period locus within and among species of the *Drosophila melanogaster* complex. Genetics 133, 375–387.

Lachaise, D., Cariou, M.L., David, J.R., Lemeunier, F., Tsacas, L., Ashburner, M., 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22, 159–225.

Lewis, P.O., 2001. Phylogenetic systematics turns over a new leaf. TREE 16, 30–37.

Maddison, D.R., Ruvolo, M., Swofford, D.L., 1992. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. Syst. Biol. 41, 111–124.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654.

Nei, M., Li, W.-H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76, 5269–5273.

Raftery, A., 1995. Hypothesis testing and model selection. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, New York, pp. 163–187.

Rosato, E.P., Alexandre, A., Barbujani, G., Costa, R., Kyriacou, C.P., 1994. Molecular polymorphism in the period gene of *Drosophila simulans*. Genetics 138, 693–707.

Rozas, J., Rozas, R., 1997. DnaSP version 3.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. 13, 307–311.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156, 879–891.

Swofford, D.L., 1993. PAUP. Sinaeur Associates.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics, second ed. Sinaeur Associates.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Templeton, A.R., 1993. The Eve hypothesis: a genetic critique and reanalysis. Am. Anthropol. 95, 51–72.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., Wilson, A.C., 1991. African populations and the evolution of human mitochondrial DNA. Science 253, 1503–1507.

Wares, J.P., Cunningham, C.W., 2001. Phylogeography and historical ecology of the North Atlantic intertidal. Evolution 55, 2455–2469.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276.

Zurovcova, M., Eanes, W.F., 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene Dhc-Yh3 of *Drosophila melanogaster* and *D. simulans*. Genetics 153, 1709–1715.